The Dissertation Committee for Summer Loomis certifies that this is the approved version of the following dissertation:

"Advanced" Arabic: Investigating Learners' Lexical Richness in the Context of an Oral Interview

Committee:

Esther Raizen, Co-Supervisor

Kristen Brustad, Co-Supervisor

Lia Plakans

Adi Raz

Mohammad Mohammad

"Advanced" Arabic: Investigating Learners' Lexical Richness in the Context of an Oral Interview

by

Summer Loomis, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2013

Dedication

To all my family members who crossed oceans of language and culture to come to America,

and especially to my grandparents

Acknowledgements

My immense gratitude goes to all my committee members. Thank you each for being instrumental in this endeavor. Thank you for all the countless hours you have spent working to improve this and for your dedication to the process. Any errors found here are of course my own, but I am thankful to have had such a wise and encouraging group of people to read drafts and provide thoughtful feedback along the way.

I would especially like to thank my dissertation committee co-chairs, Drs. Esther Raizen and Kristen Brustad, for their support, guidance, and encouragement. Like many students, I am indebted to you both and feel lucky to have worked with you for the many years that led up to the completion of this research. I could not have finished this without you both putting so much into the process.

I would like to thank Dr. Lia Plakans for offering such thoughtful support over many years and for continuing to work with me from a new department and position. It could not have been easy to juggle so many demands, but I benefited immensely from your perspective and energy, and I could not have been more fortunate to have been your student. Thank you also to Dr. Adi Raz for your input, advice, and encouragement at crucial moments. Thank you Dr. Mohammad Mohammad for your guidance on this project, as well as over the years I spent in my graduate program. I have always appreciated your willingness to entertain my questions at any time, despite their sometimes inchoate nature.

Thank you also to Dr. Mahmoud Al-Batal for advising me throughout my graduate program. I benefited immensely from your input and hours of work to help me improve my teaching and better understand this area. Thank you also to Dr. Martha

Schulte-Nafeh, Laila Familiar and Hope Fitzgerald for giving so generously of your time and energy to teaching new Arabic scholars. I would also like to thank the members of my cohort and the UT extended family who made my time so enjoyable there. There is too long of a list to include everyone, but I would be remiss if I did not thank Jung-Min Seo, Peter Glanville, Chelsea Sypher, Kevin Burnham, Greg Ebner, Martin Isleem, Charles Joukhadar, John Baskerville, Emilie and Scott Zuniga, and Alex and Melanie Magidow in particular.

A special thank you also to my transcribers, Aynur Hodge and Madeline Clark, for the hours you spent helping to transform these recordings into useable data. I could not have done this without you! Thank you also to all my students who allowed me to experiment on them with good humor, both in my teaching and research.

At the University of Washington, I would especially like to thank Dr. Terri DeYoung for encouraging me to apply to the Master's program, and for making my years there so productive and enjoyable. Without her encouragement, I doubt I would have continued in this area. My appreciation also goes to Drs. Kathy Friedman and Resat Kesaba for their friendship and support. I would also like to thank Drs. Michael Williams and Naomi Sokoloff for helping me to expand my knowledge of the Middle East in critical ways.

At Oberlin College, I would like to thank Drs. Janice Zinser, Eve Sandberg, and Ben Schiff in particular. Dr. Zinser encouraged me to see my potential as a language learner, despite my many failures leading up to that point. Without her support, I would have never thought it possible to learn Arabic, French, or any other foreign language. Drs. Sandberg and Schiff both helped me to view the world in much broader terms, and to encourage the interests that ultimately lead to this dissertation. They all helped me to

make fundamental shifts in my thinking about myself and the world, and I continue to think of each of them as role-models when working with my own students.

In my personal life, I have the deepest gratitude for my husband, for encouraging me from the beginning and for providing such constant and vital support along the way. Thank you for choosing me as your partner and for making many sacrifices along the way so I could pursue this work. Thank you also to Drs. Sam Cohn and Lynn Wallisch for your unwavering support and encouragement, and for opening your home to us and extending your warm friendship to both of us.

Thank you to my parents for constantly encouraging and supporting my interests, no matter how they started or where they might lead. Little did we know that my summer study plans would turn into this.

Finally, a special thank you to my grandparents for persevering through great depressions, world wars, and other challenges large and small. Thank you for showing us the potential of this country and what it means to be an American, irrespective of birthplace or first language.

"Advanced" Arabic: Investigating Learner's Lexical Richness in the Context of an Oral Interview

Summer Loomis, Ph.D.

The University of Texas at Austin, 2013

Co-Supervisors: Esther Raizen and Kristen Brustad

This study used recordings produced in the American Council on the Teaching of Foreign Languages' (ACTFL) Oral Proficiency Interviews (OPIs) to investigate the quantity and lexical richness of second language (L2) Arabic speakers' lexical production. The study focused on 28 full-length tests and 53 sub-samples of narration and description, selected from an initial data set of 115 OPIs. The research questions were: 1) What are the average words and words per minute (WPM) produced by Advanced-Mid rating level test takers in this data set? Do Intermediate-Mid rating level test takers produce fewer words and WPM than Advanced-Mid rating level test takers? Do Superior rating level test takers produce more words and more WPM than Advanced-Mid speakers? 2) What is the lexical variation in the Advanced-Mid samples as measured by type-token ratio (TTR)? Is this variation higher or lower than the lexical variation of test taker samples at the Intermediate-Mid and Superior rating levels? 3) How many shared words produced by learners at the Advanced rating levels are from beyond the 2,000 most frequently used words in Arabic according to Buckwalter and Parkinson's frequency dictionary (Buckwalter and Parkinson 2011)? 4) What qualitative observations can be made about test takers' narration and description attempts at the Advanced rating

levels? How do these attempts compare to narration and description attempts by test-takers at the Intermediate and Superior rating levels respectively?

The WPM and TTR values for the Advanced-Mid rating level differentiated this test taker speech from the Intermediate-Mid rating level speech. However, the WPM and TTR measures did not distinguish between the Advanced-Mid rating level and the Superior rating level test takers. In regards to word frequency, learners at the Advanced-Mid rating level did not produce shared words that were beyond the 2,000 most frequently used words in Arabic. However, the qualitative observations of the Advanced rating levels' descriptions and narrations appeared to show a difference between this group's lexical resources and those of the Intermediate and Superior rating levels. These findings and related suggestions for future research on the advanced L2 speaker of Arabic were also discussed.

Table of Contents

List of Tables

CHAPTER 1: INTRODUCTION

STATEMENT OF THE PROBLEM

This dissertation investigates the quantity and quality of vocabulary produced by advanced, non-native Arabic speakers. This work is intended to contribute to the larger agenda of investigating the connection between lexical knowledge and speaking skill for foreign language learners. My interest in this topic stems from several factors: 1) a desire to better understand the role vocabulary plays in the oral production of non-native speakers of Arabic, 2) an interest in contributing to the emerging discussion of how "advanced" language ability should be defined and how this can include learners of Arabic, and 3) a desire to contribute to the research of productive vocabulary use in a second language (L2).

Recent data from the National Middle East Resource Center's survey of students indicates that the majority are studying Arabic to reach a level in which they can function professionally (National Middle East Language Resource Center, 2011). Recent world events have led to an increased interest in reaching this level in Arabic and other languages deemed critical to U.S. national security. This interest has been encouraged by initiatives such as the National Security Education Program (NSEP), the U.S. Department of Defense's STARTALK teacher training programs, and NSEP's Foreign Language Flagship programs, the last of which funds five post-secondary programs devoted to Arabic. These represent particularly noteworthy attempts to break through what Benjamin Rifkin refers to as the glass ceiling of language achievement. Regarding learners of Russian, Rifkin reports that students typically reach an intermediate speaking level by the end of a traditional four-year university program, leaving them well below the level needed for many professional purposes (Rifkin, 2005, p. 11).

While there is much enthusiasm for reaching higher levels of skill, "advanced" speaking ability has yet to be empirically defined for Arabic L2 learners. The present research is meant to help define "advanced" for this learner population in order to bring Arabic data into the conversation with those who work on productive vocabulary use in other languages and to aid those who work with Arabic language learners specifically.

My interest in this area has grown from my multiple encounters with L2 Arabic speaking and testing. I struggled to reach the advanced level as a learner of Arabic over many years and took my first speaking test – after several years of classroom study – over the phone; my nervousness was greatly amplified by the fact that the resulting rating would help determine my suitability for a study abroad program that was critical to my learning goals. This was also my first encounter with the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI), which I will explain in more detail below. After I began teaching Arabic as a foreign language, I advised students taking speaking tests (most commonly the OPI or some version of it) and tried to better understand their experiences after they had taken these tests. My third role was as an examiner after I took a training workshop and began using these testing techniques on test takers who were not my own students.

Through all of these roles, I was struck by the ubiquitous nature of the OPI and also by the many high-stakes decisions that are made on the basis of speaking test results. As an examiner, I also noticed that the performances awarded the same rating often appeared to qualitatively differ from one another; a fact that does not necessarily invalidate the results of a test, but one that may affect face validity for test takers and users. Finally, I found that discussing "advanced" with other Arabic instructors, learners, and testers seemed to indicate the common use of markedly disparate definitions, not all of which were based upon performance. In addition, those who relied on aspects of L2

performance seemed to rely on very different components, which I was not expecting. All of these reasons led to my research interest in this area.

FOCUSING ON VOCABULARY

Al-Batal (2006) posits that acquiring extensive vocabulary may be one of the most difficult challenges facing intermediate-level Arabic learners trying to improve their Arabic language abilities. This factor may also be one that distinguishes the spoken production of advanced learners of Arabic from the production of learners at other stages of language acquisition. However, the field of Arabic language acquisition lacks research that can support or offer counterevidence for this assumption. In this dissertation, I will focus on productive lexical breadth and lexical richness as potential distinguishing factors in order to investigate whether lexical quantity and quality appear to differentiate advanced learners of Arabic from learners at the Intermediate and Superior rating levels. I am interpreting productive lexical breadth as the total number of intelligible words an L2 speaker produces in a speech sample, and using John Read's definition of lexical variation as a measure of lexical richness, i.e. the expectation that more varied vocabulary is better than repeating more of the same words (Read, 2000, p. 200). In this introduction, I will explain how I arrived at these areas of focus and the theoretical framework for this research.

DEFINING "ADVANCED" USING LANGUAGE RATING SCALES

To consider potentially useful measures of advanced speaking ability, I turn to two language rating scales, one used in the U.S. foreign language testing context and one more widely used outside the U.S. context. The ACTFL Proficiency Guidelines are well

known in the United States and are widely used to evaluate speaking samples elicited in the course of the ACTFL OPI. The ACTFL OPI is the most widely used test to evaluate speaking ability in Arabic (Eisele, 2006) and is also widely used to evaluate speaking skill in other foreign languages in secondary and university level programs in the United States (Norris & Pfeiffer, 2003). While there is debate about the type of communication it elicits and the varied ways it is used (see for example Halleck 1992; Johnson 2000; Meredith 1990), the fact that it is widely used makes it a compelling context in which to research the language produced by learners.

The ACTFL OPI is also a compelling setting in which to examine Arabic L2 lexical production because it does not specify vocabulary as a rating factor. By not including vocabulary specifications regarding quantity or quality beyond very general statements about lexical resources, the Guidelines provide no incentive for test takers to vary their vocabulary for display purposes or for examiners to consider vocabulary as a distinct rating component. In that respect, test takers should only be producing the words necessary to accomplish the test's required tasks. This provides an ideal setting in which to evaluate whether there appears a relationship between the quantity and quality of the test taker's vocabulary and his or her rating.

Further, while the ACTFL OPI may be a blunt instrument for some testing purposes, the language samples elicited at the Novice, Intermediate, and Advanced levels appear to represent useful designations for learners and instructors as they attempt to gauge language learning progress in Arabic. If we proceed with this working assumption that the samples gathered in ACTFL OPIs represent broadly differing levels of ability for Arabic L2 learners, then we can focus on how advanced test takers' vocabulary produced differs quantitatively and qualitatively from one another.

It should be noted that the ACTFL Proficiency rating scale was developed based on the U.S. government's Interagency Language Roundtable (ILR) proficiency scale (Chambless 2012 pp. 142-143). The ILR scale uses the numbers zero to five to designate levels and plus signs to indicate stronger performances at those levels. The ACTFL scale is considered to be equivalent to the lower levels of the ILR scale. For example, Elvira Swender reports the ACTFL and ILR ratings needed to function in a variety of workplaces, listing the Advanced rating levels as corresponding to 2/2+ and the Superior rating level as equivalent to 3 on the ILR scale (Swender 2003 p 525). John Eisele similarly lists the Advanced rating levels or 2/2+ on the ILR scale as the beginning of "limited working proficiency" and the Superior level or 3 on the ILR scale as "general professional proficiency" (Eisele, 2006, p. 204)[1].

---

[1] Outside of the U.S. language-testing context, the Common European Framework of Reference for Languages (CEFR), developed by the Council of Europe, is the most widely referenced scale. Unlike the ACTFL Proficiency Guidelines, the CEFR treats the topic of lexical resources as a separate component of L2 ability. While CEFR "global scale" does not use labels such as "intermediate" or "advanced" *per se*, these levels can be understood to fall into the categories of beginner (known as A1 and A2), intermediate (B1 and B2), and advanced (C1 and C2) (CEFR, 2001, p. 23).

In regard to vocabulary, the CEFR specifies that the most advanced learner at a C2 level demonstrates "a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning" (Council of Europe, 2001, p. 112). In contrast, the description for C1 incorporates similar expectations but allows for gaps, which are "…to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms" (Council of Europe, 2001, p. 112). The description's requirements of both "a very broad lexical repertoire" and "command of idiomatic expressions and colloquialisms" are of particular interest. While the CEFR deems lexical range to be a sufficiently important component of language ability to define and include, it does not define what constitutes a "good command" or how many expressions would be needed to demonstrate sufficient knowledge of "idiomatic expressions and colloquialisms."

Regarding vocabulary control, the CEFR simply states that the C2 level learner should demonstrate "consistently correct and appropriate use of vocabulary," while the C1 learner may have "occasional minor slips, but no significant vocabulary errors" (Council of Europe, 2001, p. 112). Similarly, it is up to the CEFR user to determine what constitutes "correct and appropriate" vocabulary use, but the framework does include vocabulary control, which at the very least encourages CEFR evaluators and interpreters to regard vocabulary as an important component of language ability. Some research evidence has also been found for links between CEFR levels and vocabulary size (Milton, 2010; Milton & Alexiou, 2009). As such, the CEFR may represent a viable option for use in the U.S. as this attention to vocabulary could yield important information about L2 users' abilities.

ACTFL OPI GUIDELINES

I now turn to the ACTFL Proficiency Guidelines to examine in more detail how lexical breadth and lexical variation may be implicated at different rating levels. The level descriptions are written to include both what learners can accomplish when speaking and what abilities they lack, i.e. what prevents them from receiving a rating at the next level. It should be noted that newer Guidelines were made available in February 2012 and included a level higher than the Superior rating level called Distinguished. While the research conducted in this dissertation was designed with the 1999 ACTFL Guidelines in mind, the newer version of the Guidelines provide more detailed information and will be used here. I will begin with the highest rating level included in this dissertation, the Superior level, and then discuss the descriptions of the lower levels after that.

Superior Rating Level

The first sentence in the description for the Superior level from the revised 2012 Guidelines states: "[s]peakers at the Superior level are able to communicate in the language with accuracy and fluency in order to participate fully and effectively in conversations on a variety of topics in formal and informal settings from both concrete and abstract perspectives". This description does not mention vocabulary, but has direct relevance to it in three ways. Vocabulary must be considered in evaluating: 1) the accuracy of a learner's communication; 2) the effectiveness of participation, which is based at least partially upon lexical choices; and 3) how well the learner handles a variety of topics. The description also lists the ability to participate in conversations in formal and informal settings. In the case of Arabic, this would further require an examination of

the learner's ability to use both formal and informal vocabulary, as these can differ significantly.

Superior level speakers are also expected to do the following with "ease, fluency, and accuracy":

1. Demonstrate their ability to address topics in their areas of interest and specialization;
2. Provide explanations of complex issues with requisite detail; and
3. Produce "lengthy and coherent narrations" (ACTFL, 2012, para. 4)

At this level, all of this must be done without unnatural hesitation (i.e. in searching for a word or expression) or inappropriate use of vocabulary. Similarly, Superior speakers are expected to produce "structured argument to support their opinions" and to "construct and develop hypotheses to explore alternative possibilities" when called upon to do so (ACTFL, 2012, para. 4). In theory, the speaker must have a fairly rich lexical base with which to narrate, hypothesize, and argue in detail.

Advanced and Intermediate Rating Levels

Unlike the Superior level, the Advanced level is sub-divided into Advanced-High, Advanced-Mid, and Advanced-Low. Speakers in the Advanced rating levels are expected to demonstrate an ability to narrate a story in all relevant time frames, describe, and handle a complicated situation. In addition, the ACTFL Advanced level is thought to be a necessary threshold for learners to use their language outside the classroom in a

professional capacity[2] (Judith Elaine Liskin-Gasparro, 1993). A test taker at the Advanced-High rating level should also be able to handle some of the tasks of the Superior level, including defending and supporting an opinion effectively. However, according to the Guidelines, a defining feature of Advanced-High learners should be an inability to sustain performance at the Superior level. The Guidelines state that a speaker may avoid a task or resort to concrete description to compensate for difficulty in producing language that includes hypotheses and argument.

Of interest to my consideration of vocabulary usage in Advanced OPIs, the Guidelines specify that Advanced-High learners should demonstrate an ability "to consistently explain in detail and narrate fully and accurately in all time frames." Since most learners of Arabic acquire the basic components of the past, present and future tense within their first year of study, their ability to use these correctly is contingent upon their ability to recombine these elements and produce conjugations of complex or irregular verbs. Also, of particular interest to this research is the requirement that Arabic L2 learners be able to narrate "in detail." This can be evaluated, in my opinion, only from within the context of the story or other complex utterance a test taker produces in his or her OPI. A qualitative examination of the discourse produced in Advanced level Arabic OPIs would help define language that meets this criterion and analyze the lexical production found in this speech.

In regard to lexical production, the description for Advanced-Mid states that learners at this level should possess "vocabulary [that] is fairly extensive although primarily generic in nature, except in the case of a particular area of specialization or interest" (ACTFL, 2012, para. 12). One of the challenges then is to define what

[2] There is at least one teacher certification that requires only an Intermediate-High rating level to obtain according to Krista Chambless (Chambless 2012).

constitutes "generic" vocabulary in Arabic and what constitutes more specialized vocabulary. Such a distinction would have to be made based on corpus data. In the current research, I will use Buckwalter and Parkinson's *A Frequency Dictionary of Arabic: Core Vocabulary for Learners* to begin the process of investigating which lexical items produced are more common and which are beyond the 2,000 most common Arabic words (Buckwalter & Parkinson, 2011).

For the level of Advanced-Low, test takers are understood to handle the same kinds of communicative tasks as Advanced-High and Advanced-Mid, but with less ease and fluency. The Guidelines also specify that vocabulary used at this level will "often lack specificity" (ACTFL, 2012, para. 14). Advanced-Low learners are expected to present their point of view " …with sufficient accuracy, clarity, and precision to convey their intended message without misrepresentation," but this may be accomplished through rephrasing or repeating of what has been said earlier (ACTFL, 2012, para. 15). The requirement of "sufficient" accuracy, clarity, and precision raises the question of defining this within the context of a particular speech sample.

In contrast to Advanced learners, Intermediate level learners are expected to be able to handle routine, uncomplicated tasks and social situations, and to exchange basic information related to "work, school, recreation, particular interests and areas of competence" (ACTFL, 2012, para. 17). As such, it is assumed that Intermediate level learners' vocabulary production will not convey the detail or abstraction expected at higher levels. Intermediate level learners are also expected to exhibit more hesitation and linguistic inaccuracy than Advanced level learners when trying to communicate.

FINDINGS FOR ADVANCED IN OTHER LANGUAGES

I now turn to sample findings in other languages that are pertinent to the present discussion of Arabic language learners and their lexical breadth and lexical variation. In her doctoral thesis, Judith Liskin-Gasparro (1993) focused on Intermediate-High and Advanced learners of Spanish, examining portions of 14 Intermediate-High and 22 Advanced learners' interviews in order to compare: 1) the basis of the structure of stories elicited, 2) their morphosyntactic accuracy and appropriateness, and 3) the use of communication strategies. Her results show that Advanced learners produce longer, more detailed narratives that more successfully combined narrative and descriptive elements (Judith Elaine Liskin-Gasparro, 1993).

In regard to lexical breadth, Margaret Malone focused her 1999 dissertation on the development of a simulated oral proficiency test in English. In the course of this research, Malone administered a simulated OPI test to 30 non-native speakers of English to investigate the quantity and quality of the spoken language produced by these learners. She then compared this with the language predicted by the ACTFL Proficiency Guidelines' descriptions of the different levels of speech that should be found in an OPI (Malone, 1999).

Malone focused on the quantity of language produced by speakers at different levels and determined the average words per level. Her method was in line with the assumption of the ACTFL Guidelines and training materials that learners at the Superior level would produce more words than learners at the Advanced level, and that learners at the Advanced level would produce more words than learners at the Intermediate level. Malone's findings support this assertion, namely that the mean number of words per level was higher for Advanced level speakers than that of Intermediate level speakers, and higher still for Superior level speakers when compared with Advanced level speakers.

This finding corresponds to a later finding in another oral test: Read and Nation (2006) found that the average number of words increased across the bands of the International English Language Testing System speaking test in line with rising ability, although they also found that the variation between speakers in the same level was substantial (Read & Nation, 2006)[3].

Narration

Robin provided an analysis of narration in OPIs recorded with learners of Russian (Robin, 2011). Robin gathered 54 ACTFL OPIs ranging from Intermediate-High to Superior, with 5 rated Advanced-Low, 27 Advanced-Mid, 8 Advanced-High, and 9 Superior. He transcribed and analyzed these interviews, defining narrative according to Labov's 1972 requirements and McCabe and Peterson's 1984 evidence of episodic descriptors. He notes that research on narration began by focusing largely on children's first language abilities. Bearing this in mind, Robin adapted some of the requirements of narration but recognized the inherent shortcomings of using these recordings as source material, as OPIs do not focus on the elicitation of narration *per se*. However, he argues that they provide the opportunity to examine how often narratives are volunteered, to analyze the quality of these narratives, and to compare Advanced level narratives with Superior (Robin, 2011).

Robin presents some tentative conclusions, two of which are of particular interest to the current consideration of advanced Arabic L2 learners. First, he concluded that Superior rating level test takers volunteered more and provided richer narratives than

[3] The English Language Testing System (IELTS) is an internationally standardized and administered test of English language ability. IELTS is owned by British Council, IDP: IELTS Australia, and Cambridge English Language Assessment.

Advanced rating level test takers (Robin, 2011), which supports Liskin-Gasparro's findings with learners of Spanish. This willingness to volunteer narrative and the ability to provide more detail in the process may be some of the features that distinguish between Advanced and Superior rating level speakers, although Robin acknowledges that further study is needed in both Russian and other languages.

Second, Robin notes that failed narration occurred more frequently at the Intermediate-High and Advanced-Low levels within his samples, which he interpreted as a reflection of Intermediate-High and Advanced-Low learners being less savvy in their avoidance strategies than learners at other levels (Robin, 2011). Both these points are worth exploring in Arabic language-learner data and in the context of lexical richness, as I will examine whether Advanced rating level test takers provide more detailed narratives than Intermediate rating level test takers and whether Advanced rating level test takers can narrate effectively by producing language that simultaneously advances a story while also avoiding the individual test taker's linguistic pitfalls.

Using spoken data elicited in the OPI presents several challenges, and I conclude this section with a brief overview of these difficulties. First, the test questions and format are not strictly standardized, a fact that allows examiners flexibility but also means that test takers may respond to very different prompts from one test administration to another. This also means that the length of each test varies and therefore the amount of time that each test taker is allotted to talk differs accordingly. Second, as lexical breadth and lexical variation are not specified by the ACTFL rating scale, neither examiner nor test taker need demonstrate a wide variety of lexical resources or control.

In order to address these issues, I will be using words per minute as a way to account for the different test lengths. I will also focus on requests for a specific description—that of cities—and for narration. A restricted focus on these two examiner

questions will allow me to: 1) use samples in which test takers are responding to similar requests in order to make stronger comparisons between samples, 2) use more than one task type in order to analyze a broader spectrum of learner production, and 3) focus on exchanges that are purported to be defining tasks of the Advanced level. In addition, I have chosen to focus on description and narration because these are presumed to be common examiner requests in Intermediate and Superior level tests, which will allow me to compare and contrast Advanced rating level test takers' production with production from Intermediate and Superior rating level test takers.

ADVANCED LEARNERS

I will now briefly turn to Heidi Byrnes's understanding of the goals and curricular changes that must be made to produce advanced learners. I will explain how these theoretical positions are relevant to my research, before turning to the research questions and significance of this dissertation.

In her 2005 article, Heidi Byrnes discusses less commonly taught languages (LCTL) and advanced learners, and argues for the necessity of expanding LCTL instructional goals to include reaching advanced levels in the second language under consideration (Byrnes, 2005). Rather than see this as merely a theoretical addition to the myriad other concerns within the fields of second and foreign language instruction, Byrnes argues for enshrining advanced levels of ability as a foundational principle of all language instruction and testing. In her view, this principle must guide and inform all decisions related to L2 language instruction and testing, rather than be regarded as a late addition to an already long list of existing concerns. I strongly agree with Byrnes' position in regard to the importance of aiming for advanced abilities in an L2. While

foreign language instruction is often not organized around this goal of high-level acquisition, reaching advanced levels of ability should be the fundamental organizing aspiration upon which all language instruction decisions are understood.

Like Byrnes, I recognize that making the goal of reaching advanced levels of L2 ability a theoretical and practical starting point would demand a fundamental reassessment of many entrenched instructional and assessment practices in second and foreign language instruction in the U.S. context. While daunting, this effort could produce positive effects by changing the focus of instruction away from beginning and intermediate levels to something higher and ultimately more useful.

By focusing on advanced capacities as a learning outcome, language instructors as a group would be asking how quickly and how well we can cultivate the basic skills necessary to prepare and push students into the wider arena of advanced ability. Rather than focusing on delivering particular grammar points or memorizing assigned vocabulary, we need to introduce a larger frame that supports teaching students how to learn a foreign language to a high level. This important goal can be accomplished best in my opinion by defining what the advanced level encompasses for different skills and language use areas, and then applying this knowledge to the classroom in the form of pedagogical approaches and curricular and extracurricular activities. This dissertation work will, then, take a step toward defining "advanced" speaking ability for non-native learners of Arabic.

I also agree with Byrnes' position that using "…neat distinctions about … a person's native language" to determine language ability is increasingly unhelpful, given the current globalized context in which many human beings are now called to function (Byrnes, 2005, p. 29). Instead, Byrnes sides with those who argue for understanding linguistic ability as both a product and a vehicle of social practice, and for moving away

from positing the ideal native speaker as the model of ultimate achievement[4]. I strongly agree with Byrnes that framing language ability as more socially driven and defined will necessitate fundamentally different approaches to language instruction and testing. This approach demands a wider consideration of what "knowing a language" means and how this knowledge can be presented to, practiced with, and elicited from learners. It also means considering different ways for how this ability should be measured. This question of framing is relevant to the larger context of this dissertation work, and I believe it is relevant to Arabic language instruction and testing in particular, as these areas could benefit from the employment of this theoretical starting point.

This research purposefully assumes a focus on what Vivian Cook refers to as the abilities of multi-competent, non-native speakers (Cook, 1999). Arabic language instruction in particular has suffered from an unhelpful fixation on what native speakers do and do not do, which has distracted from the more pressing goal of producing non-native speakers of the language who can interact with native speakers in a competent manner. The language produced to accomplish such a goal may differ significantly from the language produced when native speakers interact among themselves.

It is for these reasons that I have chosen the current dissertation topic. It has potential to contribute to our understanding of advanced learners of Arabic - as measured by their achievements - and to begin the work of shifting focus away from the native/non-

[4] The traditional reversion to the native/non-native speaker binary is unproductive generally and even more so when it comes to Arabic and other less commonly taught languages that are considered harder to learn. I maintain this position for three reasons: 1) the native/non-native binary instantiates a power dynamic that negatively effects native and non-native speakers alike, 2) it sets an unachievable goal for learners and undervalues their actual achievements, and 3) it deflects focus and intellectual energy from the pressing necessity to develop socio-linguistically appropriate ways of teaching and assessing L2 language use and ability.

native binary to a more useful focus on different ways we can conceive of and test advanced language production by non-native speakers of Arabic.

RESEARCH QUESTIONS

As my interest is in investigating the productive lexical breadth and lexical variation of Arabic L2 production at the Advanced level, I will use the following research questions in this study:

1. What are the average words and words per minute produced by Advanced-Mid rating level test takers in a subset of the OPIs under consideration? Do Intermediate-Mid rating level test takers produce fewer words and words per minute than Advanced-Mid rating level test takers? Do Superior rating level test takers produce more words and more words per minute than Advanced-Mid speakers?
2. What is the lexical variation in the Advanced-Mid samples as measured by type-token ratio (TTR)? Is this variation higher or lower than the lexical variation of test taker samples at the Intermediate-Mid and Superior rating levels?
3. How many shared words produced by learners at the Advanced rating levels are from beyond the 2,000 most frequently used words in Arabic according to Buckwalter and Parkinson's frequency dictionary (Buckwalter and Parkinson 2011)?
4. What qualitative observations can be made about test takers' narration and description attempts at the Advanced rating levels? How do these attempts compare to narration and description by test-takers at the Intermediate and Superior rating levels respectively?

SIGNIFICANCE

My overarching interest is in defining the kind of performance—in terms of lexical quantity and richness—that corresponds to a learner being rated Advanced. First, the purpose of this research focus is to begin building a lexical profile for advanced

learners of Arabic based on test taker performance in the ACTFL oral test. Second, the question of defining advanced ability relates very broadly to a fundamental balancing act that learners must cultivate in their spoken performance. This balance must be struck between expanding lexical production and communicative content, and minimizing the instances of miscommunication or possible miscommunication that can occur. It must be recognized that learners at the Advanced level are not able to produce error free or native-like oral language but instead must rely on their interlocutors to help them accomplish their communicative goals. The present study takes these potential limits into account and is focused on defining the advanced level for Arabic language learners by examining their lexical deployment in the context of this test.

To my knowledge, work on vocabulary production in the OPI context has not been undertaken in Arabic. This research will allow me, then, to contribute new knowledge to the fields of Arabic and applied linguistics in three ways:

1. By identifying and analyzing Advanced Arabic speakers' vocabulary generated during actual learner performance, which will lead to a better understanding of the relationship between vocabulary and Advanced speaking ability. While we may assume that a focus on lexical production will provide insight into a learner's speaking abilities in Arabic, there is no empirical evidence for that;
2. By confirming or challenging the assumption that Arabic language learners, like non-native speakers of other languages, will produce more words at the Advanced level than at the Intermediate level, and, similarly, more words at the Superior level than at the Advanced level; and
3. By documenting and providing observations of the quality of the description and narration attempts learners provide in the course of OPI interviews.

While ACTFL OPIs are widely employed, there have been no studies that focus on vocabulary production in Arabic for this test. The insights gathered from this research will also contribute to our evolving understanding of "advancedness," as coined by Heidi Byrnes across other foreign languages (Byrnes, Weger-Guntharp, & Sprang, 2006). In turn, this could have implications for curriculum design and other forms of assessment for Advanced speakers of Arabic as well as other foreign languages.

Lastly, we have witnessed a shift in language testing from focusing on theoretical rating scales to building validity arguments. Briefly stated, this shift encourages the examination of the test taker language produced in a test in light of the uses to which learners' language abilities will be put. Further, it demands that an argument be built to respond to this requirement (Chapelle, Enright, & Jamieson, 2010). In order to investigate how language elicited in the OPI would compare to language use in non-testing circumstances, we must first have a detailed understanding of the language being produced in the OPI. This research takes one of the first steps toward examining the language produced in the ACTFL OPI by Arabic language learners.

Structure of the Dissertation

This chapter has provided a general introduction to the study, including its significance and research questions. Chapter two contains a literature review of current theory and research on vocabulary testing and oral interviews. In chapter three, I focus on the methodological choices made and methods used to analyze the data. Chapter four presents the results of the quantitative descriptors and qualitative observations in the original Arabic with an accompanying English translation. Lastly, I discuss in chapter

five the findings and the limitations of the current work, and offer conclusions and future research directions.

CHAPTER 2: REVIEW OF LITERATURE

In the previous chapter, I outlined the reasons for my focus on vocabulary and the advanced learner in particular. The ACTFL Proficiency Guidelines were explained as well as a select number of findings in other L2 languages, with particular attention to narration and description. I presented my research questions for the current study and explained the potential significance of this work to Arabic L2 research and language assessment. In this chapter, I will begin by explaining the relevance of productive vocabulary and speaking tests before outlining the structure of the literature review that follows.

Paul Meara observes that vocabulary was not widely regarded as a compelling research topic in second language acquisition until the publication of Paul Nation's (1990) *Teaching and Learning Vocabulary* (Richards et al., 2009 p. xii). However, the situation has changed considerably and Norbert Schmitt (2010) now reports that the relationship between vocabulary and language ability is well-established, and that the most researched areas now explore vocabulary acquisition, word use in language, and the inherent properties of words (Schmitt, 2010, p. 3). In regards to language testing, Charles Alderson states that his work on DIALANG[5] led him to believe that the size of a test taker's vocabulary is relevant to his or her performance on *any* language test (Alderson, 2005, p. 88, emphasis added).

Rod Ellis notes the increased interest in vocabulary and vocabulary acquisition, pointing out that a quick examination of topics in the most recent issue of *Language Teaching Research* shows vocabulary to be an area of strong research focus (Ellis, 2013).

[5] DIALANG is an online diagnostic language test developed with European Union funding to assess a test taker's strengths and weaknesses in one of 14 different languages. It includes writing, reading, listening, grammar and vocabulary sections, and compares a test taker's performance to the CEFR.

Likewise, foreign language teachers continue to be interested in vocabulary acquisition. Ernesto Macaro reports 80 teachers of modern languages in the U.K. ranked both vocabulary acquisition and speaking as two of their top areas of interest for research (Macaro, 2003, pp. 3-5).

There are several reasons for this persistent attention. First, vocabulary knowledge appears to correlate strongly with general language ability in both L1 and L2 learning (see Daller & Xue 2007 p. 150; Schmitt, 2010, p. 3). For example, Alla Zareva (2005) tested 30 native and 34 L2 speakers of English, divided into intermediate and advanced learner groups. She designed her experiment to investigate which lexical measures were the most useful in detecting differences between L1 and L2 groups. She found that a verified self-report and vocabulary size were the two measures that correlated the most strongly with overall language ability and were therefore indicative of which group individuals belonged to (Zareva, 2005).

Second, it is clear that learners' lexicons grow over the time spent learning a language. James Milton tested 449 learners of French as a second language using a yes/no checklist vocabulary test in which learners reported their knowledge, and these estimates were adjusted based on learners' responses to pseudo-words included in the test (Milton, 2008). Milton's findings indicated that there was steady growth in the learners' lexicons throughout the years he gathered his data, from the participants' first year of high school through their last year at university in a British school system. He found that learners' lexicons grew steadily by about 500 words per year (Milton, 2008, p. 335). Although the precise number may not be applicable to learners of all languages, research into vocabulary growth may make it possible to set reasonable benchmarks for learners and to provide guidelines for progression in L2 learning.

Third, vocabulary knowledge appears to be an area of L2 learning in which learners may be able to acquire L1 like lexicons. Zareva studied written word association tests of native and L2 speakers of English and found results that suggest it is the quantity of learners' lexicons that differ, rather than the quality of their lexical knowledge (Zareva, 2007). Likewise, Bardel, Gudmundson, and Lindqvist examined low-frequency vocabulary in the speech of L2 speakers of French and Italian and suggest that learners may be able to cultivate native-like lexical use of low-frequency vocabulary (Bardel, Gudmundson, & Lindqvist, 2012). Andrea Hellman also speculates that the lexicon "may be the potentially most successful area of adult onset L2 learning" (Hellman, 2011, p. 162). If this is the case, then research in the productive use of vocabulary in an L2 may tell us more not only about what words L2 learners produce in response to specific prompts, but also what ultimate attainment in this area might look like for learners of various languages.

Research in receptive and productive vocabulary use has shown that these types of vocabulary learning differ, both in terms of the amount of effort needed to accomplish them (Nation, 2001, p. 32) and also in the number of words that can be used in a receptive or productive mode. Both Paul Nation and John Read have separately called for studies focusing on productive vocabulary use, arguing that we can best understand how learners employ their lexicons by investigating what words they produce when using their L2 (Nation, 2001; Read, 2000). Norbert Schmitt likewise calls the measurement of vocabulary in productive language use one of the "prominent knowledge gaps" in this area (Schmitt, 2010).

Studying L2 vocabulary use in spoken production is also an obvious choice because foreign language learners are typically very interested in developing their speaking skills. John L. Walker documented this interest in a 1973 departmental survey

of 1200 university students in which the most frequent comment received was a request for more speaking practice (Walker, 1973, p.102) . More recently, Linda Harlow and Judith Muyskens piloted a survey of 471 Spanish and French students in intermediate language courses at a U.S. university (Harlow & Muyskens, 1994). The majority of the student responses in the pilot study identified learning to speak the language as the students' number one goal (Harlow & Muyskens, 1994, p 143). Similarly, when Harlow and Muyskens used the same survey with 1,373 learners of French and Spanish in 12 universities, they found that the majority again identified learning to speak the language as their primary goal. Rivera and Matsuzawa also found that speaking was the number one foreign language learning goal identified by 48 learners of both commonly taught and some LCTLs at a U.S. undergraduate institution (Rivera & Matsuzawa, 2007).

Learners of Arabic are similar in this regard (Belnap 1987; National Middle East Language Resource Center 2011). In a 2010 survey conducted at a large university in California, Jason Vivrette found that 34 students enrolled in 2nd semester Arabic ranked, "holding a sustained conversation with an Arabic speaker" and "listen[ing] to and understand[ing] a conversation between Arabic speakers" as their two most important Arabic language study goals (Vivrette, 2010). Similarly, Husseinali (2006) investigated Arabic learner goals by surveying 120 students enrolled in first and second-year courses in a university setting. Husseinali divided his participant pool into those with Arab or Muslim religious heritage, and those without such heritage. He found that both heritage and non-heritage respondents reported the strongest levels of agreement with the goal to "converse with people" and identified this as their top learning goal (Husseinali, 2006, p. 401).

Despite this interest, few studies have focused on productive vocabulary use and L2 speaking (David, 2008). As Annabelle David puts it, "We do not have a detailed

picture of learners' access to their L2 lexicon when faced with an unprepared oral task with an unknown interlocutor, a context frequently adopted to represent communicative competence" (David, 2008, p. 317). This paucity of studies that focus on productive vocabulary use in speaking is surprising, as David noted, given the common use of unrehearsed speech as a test of learners' ability to communicate in another language and the general acknowledgement of the importance of vocabulary in language learning. Unsurprisingly, studies that have been conducted on productive vocabulary use in speaking have largely focused on learners of English or other more commonly taught languages as I will detail below, and have not yet included many samples from non-Germanic or non-Romance languages. Therefore, the addition of studies from other language families will be critical in providing support for or drawing into question findings based largely on English and other more commonly taught languages.

The literature review below is divided into three main parts: 1) vocabulary and measurement in L2 research, 2) previous studies on ACTFL OPIs, less commonly taught languages and speaking tests, and 3) previous research on L2 learners of Arabic.

VOCABULARY AND MEASUREMENT IN L2 RESEARCH

Crossley, Salsbury, McNamara and Jarvis use the broad term lexical proficiency (Crossley, Salsbury, McNamara, & Jarvis, 2011, p. 182) to refer to a speaker's ability to employ his or her lexicon, and argue that lexical proficiency should be regarded as a composite of three components of vocabulary knowledge: the number of words a person knows (lexical breadth), how well a person knows these words (lexical depth), and how quickly a word can be retrieved or processed (what Crossley and colleagues term lexical processing) (Crossley, Salsbury, McNamara, & Jarvis, 2011, p. 182). They also argue for

using spoken data to estimate speakers' lexical proficiency because of its "spontaneous and unmonitored" nature (Crossley, Salsbury, McNamara, & Jarvis, 2011, p. 189).

Using this definition of lexical proficiency, Crossley and his colleagues tested whether computerized lexical indices based on vocabulary breadth, depth of knowledge, and lexical retrieval and processing ability could predict human ratings of L2 speech in English. They used samples from 29 learners with a Korean, Arabic, Mandarin, Spanish, French, Japanese, or Turkish L1 (Crossley, Salsbury, McNamara, & Jarvis, 2011). The variables that produced the strongest correlations with human raters' evaluations of the speech samples were lexical diversity (measured using Malvern and Richard's *D*), word imagability (how easily a word can be imagined as an image), word familiarity, and hypernymy (how specific or general a word is) (Crossley, Salsbury, McNamara, & Jarvis, 2011, p. 189)[6].

Crossley and his colleagues found that lexical diversity explained 45% of the human ratings variation (Crossley, Salsbury, McNamara, & Jarvis, 2011, p. 190), and stated that greater lexical diversity in this speech data correlated with the production of less easily visualized and less familiar words. They note that this appears to differ from their findings using written samples collected from English language learners. The authors speculate that the contextual nature of speaking might allow for the production of less imagable and less familiar words (Crossley, Salsbury, McNamara, & Jarvis, 2011, p. 191). This strengthens the supposition that increased lexical diversity and use of more abstract vocabulary correspond to higher ratings of speech samples, assuming the use of a reliable rating method. In contrast, the word qualities that were among the least predictive

[6] Malvern and Richards's *D* uses a computer program that takes random samples of the words produced in a transcript and makes a best fitting curve to approximate the lexical richness of the sample (Daller and Xue pp 151-152).

were word frequency and word meaningfulness, i.e. the number of other words with which a particular term is associated (Crossley, Salsbury, McNamara, & Jarvis, 2011, p. 189). If these findings hold true for Arabic, then a reliable measure of lexical diversity should discriminate between Arabic L2 speakers at different ability levels. In contrast, word frequency counts for words that Arabic learners produce should not discriminate between speakers of differing abilities.

In a separate study, Crossley and Salsbury examined which lexical indices could predict the nouns and verbs L2 English language users produced in spontaneous speech (Crossley & Salsbury, 2010). They compiled two lists using speech corpora gathered from L1 and L2 speakers of English, and then analyzed these lists to understand which lexical indices predicted the words most frequently produced by the L2 speakers in the study. Their findings indicate that the nouns produced were "more frequent, more meaningful, and more familiar" (Crossley & Salsbury, 2010, p. 121). Verbs were also found to be more frequent, meaningful and familiar like nouns. In addition, they were found to be more general than those of the verbs not produced by the L2 speakers of English. (Crossley & Salsbury, 2010, p. 115). This lends support to the idea that lexical indices like frequency, familiarity, and meaningfulness may determine whether the words are produced or not in L2 speech.

Salsbury, Crossley, and McNamara examined the psycholinguistic information available for the lexical items produced in a spoken corpus collected from 6 learners of English over a year (Salsbury, Crossley, & McNamara, 2011). Drawing on findings from L1 studies, the authors reported that words with higher imagability and meaningfulness are considered easier to learn, while words with lower scores in these two areas are considered harder to learn (Salsbury, Crossley & McNamara, 2011, p. 356). They compared word concreteness, imagability, and word meaningfulness scores for the words

produced by their L2 speakers, using the psycholinguistic information available in the Medical Research Council Psycholinguistic Database (Salsbury, Crossley, & McNamara, 2011, p. 343). Repeated ANOVAs showed significant changes in word concreteness, imagability, and meaningfulness based on the amount of time learners had spent studying the language and that the words that learners produced became "more abstract and less context dependent" over time (Salsbury, Crossley, & McNamara, 2011, p. 343).

It is of particular interest that the concreteness scores of words produced by learners over the course of the year decreased. Salsbury and his colleagues argued that the concreteness of the words produced was unrelated to the speaking tasks used to elicit the samples. They reported that their study participants did not produce more abstract language in response to the abstract terms offered to them (such as "lonely" or "confused") until after their first month of study (Salsbury, Crossley, & McNamara, 2011, p. 355). Likewise, there were also decreases in imagability scores as learners' abilities increased, further supporting the position that learners were producing more abstract vocabulary (Salsbury, Crossley, & McNamara, 2011, p. 356). Lastly, learners produced words that had lower meaningfulness scores later in the year, indicating that the words they said had fewer other words associated with them. Although comparable psycholinguistic data is not available for Arabic, this finding suggests that learners of Arabic and other LCTLs may also produce more abstract words as they progress in their language learning.

In addition to studying word properties such as imagability and concreteness, researchers have attempted to isolate words that should be considered advanced based on considerations of both word frequency data and lexical richness measurements. Early in this type of research, it was assumed that lower frequency vocabulary would be harder to acquire, and therefore production of these words in written or spoken samples would

indicate more advanced abilities in the language. For example, Ovtcharov, Cobb, and Halter (2006) transcribed the oral production of 48 Canadian government functionaries who were classified as intermediate and advanced L2 speakers of French. Among other findings, their research indicated that the advanced L2 speakers produced a greater proportion of low-frequency words than the intermediate speakers (Ovtcharov, Cobb, & Halter, 2006, p. 115).

Lindqvist, Bardel and Gudmundson similarly experimented in using word frequency data to distinguish between speech data gathered from intermediate and advanced Swedish L1 learners of French and Italian, using native-speaker data as a control (Lindqvist, Bardel, & Gudmundson, 2011). They based their work on Batia Laufer and Paul Nation's Lexical Frequency Profiler (LFP), a measure of lexical richness that is based on lists of word frequency in English (Lindqvist, Bardel, & Gudmundson, 2011, p. 222) and used primarily on written data produced by learners of English. Although a lexical profiler is beyond the scope of the current research, the findings are useful to include, in light of the potential relationship between low frequency words and advanced speaking ability.

Lindqvist and her colleagues developed and applied an LFP for French and Italian L2 spoken data to test the measure's ability to categorize learners according to their spoken lexical production in a language other than English. In the LFP's first iterations, they developed French and Italian profilers based solely on frequency data for these languages. They found that the lexical profilers differentiated between the two groups of intermediate and advanced learners of French and Italian. However, the profilers also indicated that the scores for L2 learners of French were similar to those of the French L1 control group, while the scores for L2 learners of Italian did not indicate that they had low-frequency word scores similar to those of the Italian native-speaker controls

(Lindqvist, Bardel, & Gudmunson, 2011). This supports the position that frequency data offers some potential for differentiating learner abilities, but it is unclear why L2 learners of French would display scores similar to native-speakers' scores while L2 learners of Italian would not. In a separate article, Linqvist speculates that the uncertain results may be partially due to the fact that the frequency data was developed for written production, rather than oral production, but this did not explain the difference between the two languages (Lindqvist, 2010).

In order to refine the LFPs' abilities to place L2 learners, Bardel and Lindqvist undertook a related study. They examined the words that contributed to the lexical richness scores and noted that some thematic vocabulary and cognates were coded in the vocabulary profilers as "advanced" simply by virtue of being low-frequency words in general language use (Bardel & Lindqvist, 2011). While these words may be low-frequency in a strict sense, they may actually be common to materials used to teach L2 learners of the language or may be cognates in learners' L1 and therefore easier to acquire. If lexical richness measures are based solely on measures of frequency, then these words (which the authors refer to as "thematic vocabulary" and cognates, respectively) might drive up the lexical richness numbers and provide misleading results.

Bardel and Lindqvist therefore examined outliers in their word frequency profile data, using two learners of French and two of Italian who scored the lowest and highest respectively according to the LFP that Bardel and Lindqvist had developed based solely on frequency data. They then excluded thematic vocabulary and cognates shared between learners' L1 and L2 in order to develop a more precise profiler. Bardel and Lindqvist's findings appear to indicate that word frequency data alone may not be sufficient to distinguish between L2 learners' spoken production levels and that the definition of low-

frequency (and therefore differentiating for the purposes of lexical richness) would have to exclude some cognates and essential thematic vocabulary (Bardel & Lindqvist, 2011).

In another study focusing on lexical richness, Daller and Xue measured the lexical richness of L1 Chinese speakers of L2 English to see which measures would provide the best discrimination among differing levels of ability. They gathered samples from two groups: 26 L1 Chinese speakers studying in the U.K., and 24 L1 Chinese speakers living in China and taking English as a foreign language for their undergraduate degrees (Daller & Xue, 2007). The authors began their experiment with the assumption that longer lengths of residency in an English-speaking environment would lead to greater ability in English and that this difference would be reflected in higher lexical richness values for the group studying in the U.K. (Daller & Xue, 2007, p. 155).

Daller and Xue gave participants a picture description task and a C-test. The authors did not find a significant difference between the two groups using type-token ratio[7] (TTR) and argued that TTR does not appear to be a valid measure of lexical richness when participant ability levels differ (Daller & Xue, 2007, p. 164). However, they did find significant differences between the two ability groups in terms of the number of types (i.e. verbs, nouns, adjectives, etc.) test takers produced in the picture description task. They also found significant differences in lexical diversity. To measure lexical diversity, they used Guiraud's Index and Malvern and Richards's *D* (Daller & Xue, 2007, p. 164)[8]. Daller and Xue's study supports the position that word types and lexical diversity measures may discriminate between different groups of test takers, but

[7] TTR is a widely used measure of lexical richness, using the number of different word types divided by the total number of words in a sample. However, TTR is affected by text length, meaning that a longer text is likely to have a lower TTR than a shorter text (Read 2000 p. 201).

[8] Guiraud's Index is an attempt to overcome the limitations of the widely used Type-Token Ratio (TTR). TTR is a ratio of the different types of words over the total number of words and is effected by increasing text length. Guiraud's Index uses a square root (G = types/√tokens) in order to adjust for this limitation.

their study casts doubt on the use of TTR for measuring the lexical richness of spoken language from speakers of differing abilities.

In another study of English L2 speakers, Iwashita, Brown, McNamara, and O'Hagan analyzed 200 recordings gathered in the Test of English as a Foreign Language (TOEFL) Internet-based test (iBT) (Iwashita, Brown, McNamara, & O'Hagan, 2008). They used multiple measures on the TOEFL iBT samples to determine if vocabulary, fluency, pronunciation or grammatical accuracy and complexity distinguished between test takers at different levels. They found that while all these measures contributed to the overall ratings to a certain degree, aspects of vocabulary and fluency had the greatest impact. Specifically, Iwashita and her colleagues found that vocabulary type and token were the most important in the category of vocabulary. They also found that speech rate, defined as the total number of syllables divided by the total number of seconds and excluding pauses of three or more seconds, was the most important measure for fluency (Iwashita, Brown, McNamara, & O'Hagan, 2008, p. 34).

The preceding studies have used different means of eliciting speech samples from L2 learners and it is well known that task variation may introduce measurement error (Fulcher & Reiter 2003; Lumley & O'Sullivan 2005; Taguchi 2007). Task type may also affect the lexical diversity and lexical types found in L2 learner speech. For example, Bulté and his colleagues' longitudinal study gathered spoken data in French and Dutch using Mercer Mayer's 1969 "Frog, where are you?" (Bulté, Housen, Pierrard, & Van Daele, 2008). The participant pool consisted of 19 Dutch L1 and 19 French L1 speakers at 12-14 years of age. The French L1 speakers were used as the control group and both groups contributed samples over two years. Bulté and his colleagues used measures of types, lexical classes, and lexical diversity to measure the lexical growth of their L1 Dutch participants' speech.

They found that verbs were most common in both their L1 Dutch and L2 French samples, followed by nouns, and then adjectives; they also found that adjectives constituted a far smaller component of their data than verbs or nouns (Bulté, Housen, Peirrard, & Van Daele, 2008, p. 288). The dominance of verbs in the participants' speech data is note-worthy given Annabelle David's findings on this subject. David found that L2 French learners tended to produce more nouns than verbs in earlier stages of learning, even though they were also asked to narrate a set of pictures involving actions (David, 2008). Likewise, Ellis reports that nouns are the most likely to be learned first, and therefore it would be logical to expect that more nouns than other lexical types would be produced in lower levels of speech production (Ellis, 1995). It is possible that the different elicitation techniques in these studies may have affected the speech samples that were gathered. However, assuming the elicitation techniques were similar, the question of whether nouns or verbs make up the majority of speech production at lower levels appears not to have been definitively settled.

PREVIOUS STUDIES ON ACTFL OPIS, LESS COMMONLY TAUGHT LANGUAGES AND SPEAKING TESTS

The ACTFL OPI is widely used to evaluate speaking ability in secondary and university foreign language programs in the U.S. (Norris & Pfeiffer, 2003) and is the most widely used test to evaluate Arabic speaking ability (Eisele, 2006). The test results are also used to make high stakes decisions such as teacher certification in Arabic and other languages (Glisan, Swender, & Surface, 2013). While some may argue about its appropriateness given the type of communication it elicits and the numerous ways it is used (see for example Halleck 1992; Johnson 2000; Meredith 1990), the ACTFL OPI

continues to be a widely used test in the U.S., which makes it a compelling setting for examining L2 oral production.

The ACTFL OPI is an oral assessment administered by an ACTFL certified examiner to one test taker at a time and recorded, either face-to-face or over the phone (Swender, 2003, p. 520). According to Elvira Swender, telephone tests constituted 95% of the ratings issued in 2003 (Swender, 2003, p. 521). For the rating to be considered official, the first examiner must give an initial rating of the recording and then a second examiner must rate the same recording blindly (meaning that the second examiner does not know the first examiner's rating). If there is a discrepancy between the two independent ratings, then a third examiner settles the issue (Swender, 2003). In the words of Swender, the test is intended to appear "interactive and continuously adapt[ive]" while also following the requirements of ACTFL testing procedures (Swender, 2003, p. 520). Therefore, the test does not have a prescribed set of questions and the examiner does not follow a set script when conducting the test (Swender, 2003). As mentioned earlier, this results in tests of varying lengths and differing content. This also means that examiners can choose to introduce topics and phrase their questions to test takers in different ways.

Over the past 20 years, there have been a number of studies focused on the discourse of test taker production elicited by oral interviews (Halleck, 1995). Leo Van Lier published one of the earliest works in this area in 1989. Van Lier attempted to examine the OPI from the "inside out" and questioned whether OPIs were instances of conversation, which was an early controversial assertion made by test designers. Van Lier critically examined transcripts and recordings of OPIs, and his own experiences as an OPI test taker and examiner. He raised several objections to version of OPIs being used at the time and made suggestions for future research. Many of his suggestions were taken up by other scholars in examinations of test taker production (Lazaraton, 1992; Liskin-

Gasparro, 1996a & 1996b), examiner behavior (Brown, 2003; Halleck, 1992), rater behavior and reliability (Halleck, 1996; Thompson, 1995; Shohamy, 1983; Surface & Dierdorff, 2003), and the validity and applications of the OPI (Chambless, 2012; Dandonoli & Henning, 1990; Fulcher, 1996; Henning, 1992; Herzog, 2003; Johnson, 2000 & 2001; Kagan & Friedman, 2003; Meredith, 1990; Swender, 2003).

Elana Shohamy's early findings in 1983 support using an Oral Interview (an early forerunner of what would become the ACTFL OPI) to reliably assess L2 speaking abilities, despite what she calls its subjective nature (Shohamy, 1983). Marysia Johnson disputed initial claims made by Educational Testing Services (the testing agency that administered the OPI before ACTFL) that the OPI was similar to a conversation. Johnson instead argued that discourse analysis revealed it to be more like a research interview (Johnson, 2000). After several revisions of the OPI testing techniques and rater training methods, Swender now calls the ACTFL OPI a "valid and reliable test method" in a "conversation format" (rather than representing actual conversation) and cites Dandonoli & Henning 1990, Thompson 1995, and Surface & Dierdorff 2003 to support her position (Swender, 2003, p. 520). Irene Thompson's (1995) study tested interrater reliability across five different languages and found that interrater reliability scores were consistent across these languages, lending credence to Swender's claims of reliability. However, Glenn Fulcher takes issue with Dandonoli & Henning's 1990 study on the OPI's validity claims. Fulcher presented an argument disputing Dandonoli and Henning's methods, arguing that the resulting claims for ACTFL OPI's validity were "tenuous" at best (Fulcher, 1996).

Given this ongoing debate, there have been several calls for continued and expanded research (Chalhoub-Deville, 2003; Malone, 2003). For example, Margaret Malone argues for including learners of LCTLs because the majority of the research

conducted before her article's publication in 2003 used learners of English as participants (Malone, 2003, p. 494). Recent contributions in this area have included studies with learners of Hindi (Ilieva, 2012), Japanese (Watanabe, 2003), and Russian (Fedchak, 2007; Isurin, 2012; Kagan & Friedman, 2003; Mikhailova, 2005, 2007a, & 2007b; Rifkin, 2002, 2003, & 2005; Robin, 2011 & 2012). However, it is clear that there is still a lack of studies focused on the discourse of test takers speaking non-Romance languages in the ACTFL OPI.

Among the more recent contributions to this area, Ludmila Isurin studied the "narrative/descriptive/circumlocution patterns" in speech samples gathered from 23 monolingual speakers of Russian, 10 bilingual Russian-English speakers, and 10 speakers of Russian as a foreign language (Isurin, 2012). Isurin focused on comparing the qualities of speech samples elicited from learners and bilingual speakers to the samples elicited from native-speakers. Isurin predicted and found that native-speakers produced fewer words than the L2 and bilingual speakers (Isurin, 2012, p. 210). She also predicted and found support for the hypothesis that native-speakers' descriptions would contain fewer modifiers than the learners' and bilinguals' speech (Isurin, 2012, p. 210). Her findings imply that the educated, native-speaker standard may not be the appropriate one to apply to L2 speaker samples, given that the native speaker controls produced fewer words and fewer modifiers than both the monolingual and bilingual L2 speakers. However, she acknowledges that one of the major limitations of her study was the fact that she used OPIs to gather data from the L2 Russian learners and guided, phone interviews with the monolingual native speakers (Isurin, 2012, p. 212).

In another contribution from Russian, Kimberly Fedchak (2007) gathered six Superior level OPI recordings from L2 Russian speakers and elicited feedback on the speakers' strengths and weaknesses from five native-speaker judges, all of whom were

Moscow State University faculty. The native-speaker judges commented on over 300 individual specific phrases found in the recordings. Fedchak reports that in response to the 25 hours of interviews she collected, the judges' most positive and negative evaluations had to do with L2 speakers' word choice. Fedchak interprets the findings of her dissertation work thus: "the single most important investment that can be made in interlanguage at the Superior level is an investment in the development of a broad, diverse, richly–textured, and well–understood vocabulary" (Fedchak, 2007). While the same may not apply to the ACTFL Advanced level, it is reasonable to assume that a similar focus on vocabulary at slightly lower levels of ability may also differentiate Advanced rating level speakers from Intermediate rating level speakers.

Examiners may also play a role in the quality or quantity of vocabulary produced. Annie Brown studied examiner style variation and its potential effect on speaking test results (Brown, 2003). Although examiner and score variation are not the focus of my study, it is a possible source of variation in rating results in general, and therefore I am including these points in the overall consideration of the limitations of using speaking tests to investigate learners' lexical production. In Brown's work, she examines two tests conducted with the same candidate (referred to as "Esther"), in the context of the International English Language Testing Service Speaking Module requirements. Brown chose two interviewers, one male and one female, to conduct tests with Esther to compare interviewer style and its effect on the language elicited throughout the interview. Brown did not examine gender as a factor in her study, citing a study by O'Loughlin (2000) that found no evidence for the sex of the interviewer affecting IELTS ratings nor any support for the commonly assumed differences between female and male interviewer speech.

Brown used data from a previous study by Brown and Hill 1998. In this study, she and her co-author gathered multiple speaking tests with a single subject (Esther). Brown

then chose two interviews to use as stimuli, one given by the examiner ranked as the easiest (Pam) and one by the examiner ranked as the hardest (Ian). Brown then elicited eight independent ratings of the interviews with Esther, ensuring that no rater listened to both Pam and Ian's recordings during the course of any one listening session. The independent examiners gave Esther a higher mean score for her interview with Pam than they gave for her interview with Ian.

Brown used conversation analysis to analyze both examiners' interviewing styles and argued that the differing strategies employed by the examiners resulted in qualitatively different speech samples, despite the fact that they came from the same test taker, were elicited under the same testing conditions, and were recorded on the same day. Specifically, Brown argued that the examiners' differing approaches to dealing with topics contributed to the variation. Pam appeared to use closed questions to introduce topics ("Do you have a room on your own?") and then followed up with a broader question to elicit description ("Can you describe your room to me?") (Brown, 2003, p. 9). Brown also noted that Pam engaged in topic recycling, maintained topic continuity between her prompts, and consistently closed topics in a similar fashion (Brown, 2003, p. 10-11). Ian also began with closed questions, but he did not follow up as Pam did when Esther did not elaborate on her response. His questioning style appeared less explicit to Brown and appeared to elicit much less language than Pam's questions (Brown, 2003, p. 13).

Lorenzo-Dus and Meara examined all of these areas at once by looking at test taker word types and variation, and examiner accommodation and ratings of vocabulary use in a performance test. They gathered samples from 29 high school L1 English students taking an oral test in Spanish. They recorded the tests conducted by an examiner who was unknown to the students before she administered the test (Lorenzo-Dus &

Meara, 2005). The examiner was asked to rate the students' use of vocabulary as well as other qualities of their speech, including pronunciation and fluency (Lorenzo-Dus & Meara, 2005, p. 244). The authors' working hypotheses were as follows: 1) that both the total number of word types and diversity values would correlate with the vocabulary grades, and 2) that low levels of test taker vocabulary production would correspond with higher levels of examiner accommodation. The authors examined the interviews for instances of examiner accommodation that had the greatest impact on test taker vocabulary production and identified three strategies: 1) when the examiner simplified questions or statements, 2) when she supplied or completed missing vocabulary, or 3) when she used confirmation questions to check her own or the test taker's understanding (Lorenzo-Dus & Meara, 2005, p. 248).

In addition to examiner strategies, Lorenzo-Dus and Meara found that the total number of word types was significant and discriminated between ability levels (Lorenzo-Dus & Meara, 2005, p. 245). They found that there was more vocabulary variation among the students who received higher vocabulary grades of As and Bs than among the students who received lower vocabulary grades of Cs and Ds. They interpret this as an indication that the examiner took lexical diversity into account when asked to assess L2 learners' vocabulary production. Second, they found that increased examiner accommodation correlated with lower grades on the test (Lorenzo-Dus & Meara, 2005, p. 248). This provides support for what is observed anecdotally in speaking tests, namely that examiners appear to rephrase and intervene more when a test taker has difficulty responding. However, Lorenzo-Dus and Meara found that lexical diversity (as measured by Malvern and Richards's *D*) did not discriminate between abilities in their test taker groups (Lorenzo-Dus & Meara, 2005, p. 245). This raises the question as to whether or

not these test takers were too similar in ability for lexical diversity measures to discriminate between them.

In addition to examiner effects, the nature of the speaking task may affect the language produced; for example narration and descriptions may require different vocabulary. Judith Liskin-Gasparro published a case study of one L2 speaker of Spanish who – by coincidence – narrated the same story twice in two different ACTFL OPI tests administered at the beginning and end of a summer program, one test resulting in an ACTFL Intermediate[9] level rating and the second resulting in an ACTFL Advanced level rating (Liskin-Gasparro, 1996a). Liskin-Gasparro transcribed and analyzed both recordings, and found that both her subject's stories met the requirements for narration set forth by William Labov (Labov, 1972) and both had the same content. However, Liskin-Gasparro found that the subject's story at the ACTFL Advanced level showed an ability to address the listener's perspective and better integration than it had at the ACTFL Intermediate level (Liskin-Gasparro, 1996a, p. 183).

In contrast, Julia Mikhailova focused on the task of description in OPIs and in Simulated Oral Proficiency Interviews (SOPIs), the latter using the same rating scale with recorded prompts in place of live testers (Mikhailova, 2007a). Mikhailova chose description as her task focus because it is a defining ability of the ACTFL Advanced rating band, according to Elvira Swender's 1999 ACTFL OPI tester trainer manual (Mikhailova, 2007a, p. 588). The ACTFL OPI tester manual differentiates between description and narration, stating that description should focus on a place, object, or person, and narration should revolve around an event or events (Swender, 1999, p. 122). Mikhailova compared the tasks that were meant to elicit description in 31 OPIs and 38

[9] The ACTFL Proficiency Guidelines did not include sublevels until the 1999 version (Surface & Dierdorff 2003 p. 508).

SOPIs collected in Russian (Mikhailova, 2007a, p. 588). She found that the SOPI prompts on how people's birthdays are celebrated and how test takers spend their summers gathered narrative responses instead of the intended descriptions (Mikhailova, 2007a, p. 588-590). The most successful prompts in both OPIs and SOPIs were requests for descriptions of places or cities, rather than people (Mikhailova, 2007a, p. 594).

Previous work on the ACTFL OPI, LCTLs, and speaking tests can be characterized as covering three broad areas: 1) characteristics of the test in question and the performances produced in response to a particular test, 2) aspects of examiner behavior and test taker performance, and 3) test tasks. With regard to the ACTFL OPI specifically, there was some early debate surrounding its characterization as a conversation, examiner reliability, and test taker performance. The issues surrounding labeling it a conversation have largely been settled; it is not a conversation in the ways in which applied linguists conceive of conversation. Elvira Swender, for example, characterizes the OPI as taking place in a "conversation format," by which she appears to mean simply that there are two people talking to one another during the test. However, the research findings are less clear in regard to examiner behavior in the OPI and other speaking tests, and aspects of test taker performance like vocabulary use. What has been established is that examiner behavior affects test taker performance, which appears to correlate with the vocabulary produced by test takers. Test tasks may also affect the vocabulary produced by test takers, and it appears that requests for narration and description, particularly in the case of the OPI and SOPI, may sometimes be difficult to distinguish from one another. Although the number of studies focusing on description is limited, Julia Mikhailova's work appears to indicate that descriptions of cities elicit the desired test taker speech better than requests for descriptions of other topics.

PREVIOUS RESEARCH ON L2 LEARNERS OF ARABIC

Mahmoud Al-Batal notes that the interest in L2 Arabic speaking began in the 1970s and 1980s in response to broader changes in foreign language education (Al-Batal, 1995, p. 116). One of the earlier studies in this area, Ahmed Fakhri's case study (1984) involved the elicitation of speech samples from a L2 Arabic speaker over the course of four weeks. His study subject had spent three years in Morocco; however, when he gathered speech samples from the subject, she had not used Moroccan Arabic for four years. Using these samples, he argues that communicative strategies are not used at random, but rather that this L2 speaker of Arabic used specific communicative strategies in particular ways in order to further narratives she would otherwise not have been able to maintain.

As mentioned in the introduction, Al-Batal has argued for regarding vocabulary learning as a core need for L2 learners of Arabic at all stages of learning. To that end, he has called for more research into several areas, including word frequency counts for Arabic words and investigating the ways in which Arabic L2 learners' linguistic skills are affected by their lexical resources (Al-Batal, 2006, p. 339). Salim Khaldieh made a significant contribution in this area by studying the effects of knowledge of vocabulary and case markings on L2 Arabic learners' reading comprehension. In a study conducted with 46 participants, he found that vocabulary scores had a high correlation (Pearson's $r = .90$ with $p < .001$) with reading comprehension scores, but that knowledge of the case marking system did not (Khaldieh, 2001).

Given that Arabic is a Semitic language, some research in this area has also focused on word patterns and roots as an estimate of Arabic vocabulary. For example, Sami Boudelaa and William Marslen-Wilson found that there are 2,324 word patterns

and 5,336 root combinations currently in use in Modern Standard Arabic (Boudelaa & Marslen-Wilson, 2010). This finding indicates that the challenge before learners is formidable, even if they do not need to learn every possible word or root combination. Giselle Khoury's recent dissertation (2008) tested the hypothesis that instruction in the root and pattern systems in Arabic would lead to increased morphological awareness, which would lead subsequently to increased vocabulary acquisition. Using immediate and delayed post-tests on 109 non-native learners of Arabic in their first or second semester of instruction, Khoury found a facilitating relationship of root and pattern instruction, and an increased ability to interpret unfamiliar words. However, Khoury did not find any effect on learners' retention of new words despite the fact that root and pattern instruction allowed them to make educated guesses about meaning (Khoury, 2008).

Other notable contributions include May George's dissertation (2011) on teacher scaffolding and novice L2 Arabic learners and Nader Morkus's (2009) dissertation examining the pragmatic competence of intermediate and advanced L2 Arabic speakers. However, there are no studies to date that focus on spoken vocabulary production and advanced L2 speakers of Arabic that I am aware of; this may be attributable to the fact that fewer learners reach the advanced stage of Arabic via classroom instruction compared with learners of other languages. Brown reports that, according to the Modern Language Association's report of 2006, the ratio between first and second year students of Arabic and those enrolled beyond the second year level was 8:1 (Brown, 2009, p. 407). This is obviously worrying to those interested in LCTLs, as this includes the enrollment increases produced by events of the early 2000's. This ratio compares unfavorably with Spanish, which has only five 1st or 2nd year students to each upper level student, and Chinese, which reports a slightly better ratio of 9:2.

There are notable programs working to remedy this situation. For example, the Center for Arabic Study Abroad (CASA) at the American University in Cairo is known as "a model of best-practice in Arabic study abroad programs" (Ryding, 2006, p. 17). CASA has been training advanced L2 learners of Arabic since 1967 (Abdalla, 2006, p. 321) and most recently welcomed 46 more fellows in 2012 (Soliman, 2012, p. 4). In addition, the previously mentioned NSEP Foreign Language Flagship programs have begun establishing themselves in this regard as well. The NSEP Flagship program annual report notes that four Arabic learners scored at the ACTFL Superior rating level, after finishing the Overseas Flagship program (National Security Education Program, 2012, p. 40). However, the number of advanced L2 learners of Arabic still remains small in comparison to learners of other languages in the U.S. context.

These smaller numbers also have implications for language testing. For example, Winke and Aquil note that there were not enough advanced students among the 500 recruited to test the validity of the Superior-level questions for the Online Arabic Proficiency Test developed by the Center for Applied Linguistics in 2000 (Winke & Aquil, 2006). Unfortunately, this confirms what many Arabic instructors and learners observe anecdotally in programs with more than one level, i.e. that many learners stop taking Arabic in the university classroom after the first or second year. This means that the small number of learners who choose to study Arabic typically shrinks to an even smaller number at the advanced level.

Examining commonly held stereotypes about Arabic language study may help shed light on why so few learners reach higher levels of ability. Bergman's 2009 article includes some of the myths that surround Arabic language study and teaching. She documented her experiences working as a professor of Arabic and serving as the executive head of the American Association of Teachers of Arabic beginning in 2007.

She noted the increased enrollments after September 11, 2001 and argued that the field of Arabic instruction is now entering an important shift toward longer-term planning (Bergman, 2009, p. 2).

Although these developments are encouraging, Bergman identified several points about Arabic that she believes need to be addressed to encourage better language instruction and learning. Although Bergman's observations are based on her own personal experience, I present them here because they are relevant to the research I am undertaking. Of the myths Bergman identified, the most well-known is that Arabic is an impossible language to learn. Bergman noted that both native and non-native-speakers have reached high levels of ability in the language, and proposed that this myth be reframed as "Arabic is not inherently different; it is simply time-consuming" (Bergman, 2009, p. 3). While this quotation may appear flippant taken out of context, the tone of Bergman's writing suggests this is intended in a very considered and serious sense. She cited the often-referenced contact hours required for learners of Arabic (and Chinese, Japanese, and Korean) to achieve ILR 3/3+ and argued that this represents a challenge, but in her opinion one measured in terms of time required, rather than inherent difficulty.

I agree with Bergman that many non-native-speakers can and do reach high levels of ability in Arabic, but it is difficult to point to documented, published achievements that are not produced by a particular program (for a notable exception of several L2 learners who acquired Arabic to a high level see Ioup, Boustagui, El Tigi, & Moselle, 1994). This is one way the current research will contribute: by providing vocabulary-related data to understand what can be considered "advanced" ability for L2 Arabic speakers' vocabulary. In this study, the backgrounds of the individuals and their previous study history are unknown; they will be judged on their productive lexical ability, rather than how or where they learned to produce this language.

There is also a clear need for more studies of productive vocabulary use, particularly in less commonly taught languages (LCTLs). However, there are several challenges to researching productive vocabulary use in speaking, and in Arabic in particular. As Annie Brown's findings attest, it is important to acknowledge examiner style and examiner accommodation as a source of potential variation. While I will not be addressing this in the dissertation directly, I will offer some observations on this point in the Arabic interviews I examine.

Given the findings from lexical richness in other languages, the following are my working hypotheses for this research: 1) ACTFL Advanced rating level participants' token production throughout a test should be greater than the production of ACTFL Intermediate rating test takers and lower than the production of the Superior rating test takers, 2) higher lexical richness should distinguish higher test ratings, and 3) task type may affect lexical richness, so description and narration samples must be isolated to compare the lexical richness and lexical breadth produced in response to these prompts.

Hypothesis #1 will be operationalized as average words and average words/minute produced over the length of the entire test. Given that examiners are permitted to conduct tests of various lengths by the ACTFL testing procedures, the number of words will be divided by the number of minutes recorded in each sample to account for this variation. The average number of words per minute is expected to be lower for Intermediate rating level test takers, higher for Advanced rating level test takers, and highest for Superior rating level test takers. Independent-samples t-tests will be used to discern whether or not this difference is significant.

For hypothesis #2, lexical richness will be measured by TTR. Although there is some controversy about the suitability of using TTR to measure lexical richness in spoken data, this measure will be taken to see if there are any variations worth noting

between students with different ratings. Although lexical richness is assumed to differentiate between some groups of L2 speakers, TTR measures are not specifically expected to differentiate between ACTFL Intermediate and ACTFL Advanced test takers in this study, given the potential variation among samples. However, TTR measures are expected to differentiate Superior rating level speakers from the Intermediate and Advanced rating level test takers. Word frequency beyond the 2,000 most common words will be noted for the words produced at different rating levels, but it is not expected to correlate with rating level received. Instead, it will provide another descriptor of the vocabulary produced by test takers.

Hypothesis #3 cannot be fully supported or refuted using the present data, but I will calculate words per minute and TTRs for sub-samples to see if there are any suggested patterns. Finally, while the current research will focus more on quantitative measures, I will also provide qualitative observations of the description and narration samples elicited at different rating levels. Having discussed the research findings and theoretical assumptions that support this study, I will detail the methods chosen to analyze the data in chapter 3.

CHAPTER 3: RESEARCH METHODS

In the previous chapter, I discussed some of the major findings in vocabulary research that apply to the current research. Second language vocabulary learning may be one of the few areas where learners can achieve native-like competence. However, many factors affect the vocabulary learners acquire and use productively. Among them, word imagability and concreteness have been proven to facilitate vocabulary learning. Lexical richness and lexical breadth also appear to distinguish more advanced speakers from less advanced speakers. Other factors such as word type and frequency ranking may predict the words L2 learners use, but the findings in these areas are less clear. I also briefly discussed previous research that has been conducted on the ACTFL OPI and on L2 learners of Arabic, and the hypotheses of the current study.

In this chapter, I describe the methods used to address my research questions. The goal of this study was to investigate Advanced L2 Arabic speakers' lexical production in terms of both its quantity and diversity. The data set consisted of Arabic L2 speakers' responses to questions posed by ACTFL OPI examiners in the course of recorded tests. I operationalized the concept of lexical breadth as the number of words and average words per minute produced by test takers at the Superior, Advanced-Mid, and Intermediate-Mid[10] levels in order to document these ranges for L2 speakers of Arabic according to rating level. Lexical richness was measured using TTR[11]. I also investigated lexical frequency by determining the number of words test takers produced that were not among

[10] See p. 5 below in Methodology Used for Lexical Breadth for an explanation of why I chose the "Mid" sublevels.

[11] As I noted in the literature review, TTR has been a widely used in L2 vocabulary research, but different studies have drawn different conclusions about its accuracy in measuring lexical variation. I chose TTR to provide some baseline numbers for Arabic L2 research, but recognize that using other measures of lexical variation is needed in the future to allow for comparison between L2 Arabic and L2 speaker variation as measured by other measures like Malvern and Richards's *D*.

the 2,000 most commonly used words based on work by Timothy Buckwalter and Dilworth Parkinson (2011). Lastly, I provided observations from a qualitative analysis of test takers' description and narration attempts across different rating levels by considering the content of their responses and the ways in which examiners appeared to react to test taker language.

DATA SOURCES AND LIMITATIONS

ACTFL provided 179 recordings of official, double-rated OPIs in Arabic at different rating levels. From this original data set, I excluded any recording in which the test taker's name appeared to be of Arab origin, the test taker self-identified as a native or heritage speaker, or reported being born in or spending time as a child in an Arabic-speaking country. This provided a sufficient number of samples at all ACTFL rating levels except the Superior rating level. There were only five recordings of speakers who appeared to be non-native, non-heritage speakers in the data set at the Superior rating level. Because of the relatively small number of recordings at this level, I obtained three additional Superior rating level, single-rated samples from an OPI tester. Table 3.1 shows the interviews used in this study, after this elimination process, categorized by ACTFL rating level.

Table 3.1: Total number of ACTFL OPI interviews in the data set

Superior	5 double-rated +3 single-rated recordings
Advanced-High	11
Advanced-Mid	17
Advanced-Low	18
Intermediate-High	18
Intermediate-Mid	20
Intermediate-Low	23
Total	115

All samples were produced under normal testing conditions and double-rated, with the exception of the three single-rated Superior level recordings added to the data pool later. All of the latter tests were conducted in 2011 under normal testing conditions, again, except for the four single-rated tests. However, I do not know if these tests were randomly selected from the available recordings or what fraction of Arabic-language testing recordings they represent for 2011. Additionally, I retained the recordings of some test takers whose L1 may not have been English; therefore, the sample may have included some variation due to the fact that the test taker's L1 may have affected his or her spoken production in Arabic. Similarly, there were some individual variations in the form of stuttering (one test taker self-identified as a stutterer and two others appeared to have stutters) and volubility, which may also be seen as limitations.

Although examiner variation is not the focus of the current study, the variation examiners can introduce must be acknowledged. The data set included nine different examiners; I determined this either by noting an examiner's name before the start of a test or by comparing voices between recordings. The first examiner encountered in the recordings is referred to as E1 throughout this study and all subsequent examiners are referred to by the order in which they appeared in the data set. E1 and E2 performed approximately 60% of the tests. The remaining tests were largely performed by E9 (approximately 16%), E8 (about 10%), and E7 (less than 5%). Table 3.2 shows a detailed distribution across examiner and interview rating level.

Table 3.2: Number of interviews by Examiner and Rating Level

Examiner Code	Number of Interviews Conducted per rating level and % of total interviews
E1	1 Superior (S); 4 Advanced-High (A-H); 4 Advanced-Mid (A-M); 4 Advanced-Low (A-L); 5 Intermediate-High (I-H); 9 Intermediate-Mid (I-M); 12 Intermediate-Low (I-L) 39/116 = 33.6%
E2	3S; 1A-H; 3A-M; 7A-L; 8I-H; 6I-M; 3I-L 31/116 = 26.7%
E3	0S; 1A-H; 1A-M; 1A-L; 0I-H; 0I-M; 3I-L 6/116 = 5.1%
E4	0S; 3A-H; 0A-M; 0A-L; 0I-H; 0I-M; 0I-L 3/116 = 2.5%
E5	0S; 0A-H; 0A-M; 0A-L; 0I-H; 1I-M; 0I-L 1/116 = .08%
E6	0S; 0A-H; 0A-M; 0A-L; 0I-H; 0I-M; 1I-L 1/116 = .08%
E7	1S; 0A-H; 0A-M; 0A-L; 1I-H; 1I-M; 2I-L 5/116 = 4.3%
E8	0S; 2A-H; 4A-M; 2A-L; 1I-H; 2I-M; 0I-L 12/116 = 10.3%
E9	3S (single-rated); 0A-H; 5A-M; 4A-L; 3I-H; 1I-M; 2I-L 19/116 = 16.3%

Examiners appear to have stable styles across interviews (see Ross, 1996 for example), and therefore the restricted number of examiners, while not part of the research design *per se*, may be viewed as diminishing potential variance between recordings. Additionally, Table 3.2 above shows that E1 and E2 conducted interviews at every ACTFL rating level, which would further diminish some of the inherent variation in the data set, assuming these examiners maintained consistent styles of questioning in the tests they conducted.

METHODOLOGY USED FOR LEXICAL BREADTH

In order to generate an average of word production across levels, I transcribed tests from eight Superior, ten of the seventeen Advanced-Mid, and ten of the twenty Intermediate-Mid rating level samples to calculate an average for the number of words produced by each test taker throughout the test, from the warm-up to the end of each recording. I chose the "Mid" sub-levels to base my measurements on because a test taker who receives a "mid" sub-level rating should produce a performance that is solidly reflective of the level required, according to the ACTFL rating scale. In other words, the test taker who receives a "Mid" sub-rating level should neither be struggling enough to warrant a rating of "low" nor demonstrating sufficient ability expected at the next level to receive a rating of "high"; the test taker's performance should be indicative of what this level is expected to produce and therefore should yield more representative lexical breadth measurements.

I chose ten from the Advanced-Mid and Intermediate-Mid levels each to keep the number of samples comparable to the nine Superior rating level recordings. In the rating

levels where the number of samples exceeded the number I needed, I used an online random number generator[12] to select the files I would transcribe. After using the random online number generator, the group of tests chosen included at least one test conducted by E1, E2, E7, E8, and E9, with the majority conducted by E1 and E2. This is unsurprising given the fact that these five examiners conducted approximately 85% of the tests in the data sample, but it does confirm that there are a similarly limited number of examiners represented among the recordings used to produce the lexical breadth averages.

PILOTING THE TRANSCRIPTION METHOD

Initially, I piloted transcription methods on a restricted number of recordings and then used the resulting methods on the entire data set. During the piloting phase of transcribing the samples, I listened to the entire recording of several samples from each rating level to determine the topics discussed and note where description and narration requests appeared, in order to transcribe appropriate portions later. For this piloting phase only, I excluded Intermediate-Low and Superior rating samples. In the Intermediate-Low rating level, there was only one request that could be vaguely interpreted as meant to elicit narration. However, it was neither direct (i.e. the question asked for an event the test taker remembered from her time abroad and did not include the words "narrate" or "tell me" or "story") nor successful (i.e. the test taker did not understand the question and did not produce a story in response). In regard to descriptions, there were nine requests for descriptions of various topics at the Intermediate-Low rating level. However, test

[12] http://www.random.org/integer-sets/ accessed on June 7, 2012.

taker responses to these requests were very short and they appeared not to understand the questions often, so I excluded these samples from the piloting phase.

The Superior rating tests were also excluded because there were no requests for narration among the Superior rating level recordings in the data set and because I was primarily concentrating on the Advanced rating levels. After this elimination process, I chose to concentrate in the piloting phase on samples from the Intermediate-Mid, Intermediate-High, and Advanced-High sub-levels. I chose the Intermediate-High and Advanced-High rating level samples because these should represent some of the strongest test taker responses according to the ACTFL rating scale; I worked with five Intermediate-Mid, nine Intermediate-High, and eight Advanced-High samples.

While piloting the transcription methods, I also listened to all 115 interviews in order to immerse myself in the data and to record each topic discussed by examiners and test takers. During these listening sessions, I compiled a master list of the topics covered in each recording, including all requests for description or narration. After finishing the piloting phase, I applied the resulting transcription methods to all of the subsequent transcripts produced in the course of this study.

I also used the resulting city description transcripts to compile by hand a list of the vocabulary used within samples at the Intermediate and Advanced rating levels. I focused on the description sub-samples in this phase because I assumed that the shared topic of city descriptions would increase the likelihood that test takers produced some of the same words. In order to compile this list for the description attempts, I created a Microsoft Excel spreadsheet and did the following:

- started with the lowest rating level in order to understand how the higher-level samples added to the vocabulary pool

- entered each word in the order it appeared in the test taker's description
- used each subsequent description to mark words as previously produced in other samples or as new contributions to the overall word pool
- counted the total number of times each word was mentioned across the descriptions at both the Intermediate and Advanced levels

This allowed me to examine the descriptions more closely at the word level.

Producing the Lexical Breadth Estimates

The lexical breadth estimates were generated to answer research question #1: Do Advanced-Mid rating level test takers produce more words and more words per minute than test takers who receive Intermediate-Mid rating levels? Also, do Advanced-Mid rating level test takers produce fewer words per minute than Superior rating level test takers? I produced the lexical breadth estimates using a portion of the data set as shown in table 3.3.

Table 3.3: Data set used to produce lexical breadth estimates

Superior	8 of 8
Advanced-Mid	10 randomly chosen out of 17
Intermediate-Mid	10 randomly chosen out of 20
Total	28

Transcriptions of the ten Intermediate-Mid, ten Advanced-Mid and eight Superior rating level samples included every complete word, but not sighing or overlapping speech between examiner and test taker unless the test taker's contribution was comprehensible.

Two other transcribers helped with the transcription, following my method. I reviewed and finalized all transcripts produced in the study before using them to generate the lexical breadth and lexical richness estimates presented in Chapter 4. To prepare the transcripts for the word count, I did the following: 1) inserted a space between "and" (و) and "thus" or "so" (ف)[13], 2) eliminated any repetition from the transcripts where a test taker restated the same word or phrase immediately after its initial utterance, 3) eliminated any transcription codes that indicated an incomprehensible word or words, and 4) eliminated all examiner speech. In regards to number one, I separated instances of the words "and" and "so" from the words that followed them in order to more easily review the files that WordSmith Tools would create. However, I subtracted these from the total number of words for each transcript so that "and" and "so" were not counted as separate words.

Once the transcriptions had been reviewed and prepared, I used WordSmith Tools 6.0 to produce a WordList file for each transcript. I reviewed these word lists to find typing errors and typing inconsistencies that would cause WordSmith Tools to categorize words as different from one another when they were in fact the same. The most common typing inconsistency had to do with words that had the letter hamza in them or used marks like a doubled-fatha. For example, WordSmith Tools categorized أيضاً and أيضا as different words because the second lacks the doubled-fatha marking at the end. Therefore, I had to unify the typing of all these words so WordSmith Tools would correctly group all uses of a word with a hamza or other markings together and not count them as separate types.

[13] In Arabic, a person writes "and" and "thus" without a space between them and the word that follows them. When I prepared my transcripts, I followed Arabic convention in this regard.

In my review of the Superior level test taker production, I found that one test taker's word count was far larger than the other samples at this level, even though his recording was not the longest. Upon reading the transcript again, I noticed that this test taker used the interjection "like, you know" or "that is" (يعني) very frequently. After counting these, I found that this test taker was an outlier who used this filler 296 times over a test of 25 minutes and 29 seconds. Since this word was overwhelmingly used to fill pauses rather than to provide explanation, I counted it separately in all the other test taker transcripts and deducted the count from the overall word total.

I also calculated the TTR for the transcripts I used to generate lexical breadth estimates. As has been stated previously, these test recordings vary in length. TTR is sensitive to text length because it is a ratio of word types to tokens. This means that test takers able to speak for longer may drive down their TTR by producing more tokens than test takers whose responses are shorter and include fewer words. Therefore, I used the total number of words per transcript to determine the token sample size I would use in order to calculate a TTR based on the same number of words. I found that using 500 tokens allowed me to include all but one of the full-length transcripts in my data set so I chose this as the threshold for calculating the TTR. I used the first 500 tokens from each transcript and generated a WordList file in WordSmith Tools, which produced a TTR for each file based on the same number of tokens. My transcripts were all in Microsoft Word initially and I had to prepare them as plain text files. In this process, I found that the word counts in Microsoft Word and WordSmith Tools varied slightly from one another so I included a sufficient number of words to ensure that WordSmith Tools counted 500 tokens for each sample.

In order to isolate the description and narration sub-samples, I reviewed the entire data set again and marked all the instances of description or narration that needed to be

transcribed. Table 3.4 shows the total number of tests in which narration and/or description samples were found, separated by ACTFL rating level.

Table 3.4: Total number of tests in which narration and/or description was found

Superior	4 out of 8 (with 0 tests providing samples in both categories)
Advanced-High	2 out of 11 (with 1 test providing samples in both categories)
Advanced-Mid	7 out of 17 (with 1 test providing samples in both categories)
Advanced-Low	9 out of 18 (with 1 test providing samples in both categories)
Intermediate-High	13 out of 18 (with 2 tests providing samples in both categories)
Intermediate-Mid	11 out of 20 (with 4 tests providing samples in both categories)
Intermediate-Low	1 out of 23 (with 0 tests providing samples in both categories)
Total	47

In the sub-sample transcripts of description and narration, I finalized all the transcripts and counted the number of shared words produced by all Intermediate, Advanced, and Superior rating level test takers. I then determined which of the words used by half or more of each test taker group were not among the 2,000 most commonly used words in Buckwalter and Parkinson's frequency dictionary by hand (Buckwalter & Parkinson, 2011). I also calculated the TTR for both description and narration sub-samples. I found that 100 token sample size allowed me to include the most samples from the data sub-set, so I calculated the TTR based on the first 100 words of each sub-sample.

I also used WordSmith Tools version 6.0[14] to compile lemmatized[15] lists of the content words that test takers used most frequently in their descriptions and narrations. When I consolidated words for a particular sub-sample, I did the following: 1) eliminated any English words, 2) deleted words that were incomprehensible or were corrected by the test taker by using another word, 3) combined two words that were the name of a place like "New Jersey", 4) combined definite and indefinite uses of the same word, 5) combined singular and plural uses of the same word, and 5) combined uses of words that had possessive or subject pronouns attached to them (like ولكن and ولكنهم). I compiled one list for each sub-sample transcript gathered for description or narration[16] in order to check if my hand processing was accurate.

NARRATION

I understand the evaluation of communicative ability, which stands at the core of the OPI, as a holistic assessment of a test taker's ability to respond to a particular communicative request. This approach has led me to focus on how well Advanced rating level test takers are able to narrate, despite the errors that are naturally produced in the course of their narration attempts. In identifying narration attempts, I drew on William Labov's definition of narration as a form of speech that contains events that are "reportable," and may also include a précis of what is to follow; some orientation toward

[14] WordSmith Tools is a computer program package developed by Mike Scott at the University of Liverpool. It can be used to create concordances, word lists, and key word lists from user created corpora in several languages, including Arabic.

[15] Lemmatizing means combining different inflections of the same word into one word entry, similar to what is found in a dictionary where words like "to walk," "walked," and "walking" are grouped together.

[16] In using this method I followed John Read and Paul Nation's work compiling core vocabulary lists per task. I should note here that while I used test segments, Read and Nation compiled their lists across entire tests, using a corpus of speaking tests for the International English Language Testing System (Read & Nation 2006).

the time, setting, or actors involved; and an evaluation of the story or its components (Labov, 1972, p. 370). I follow Labov in defining *minimal narration* as the production of a single, temporally related juncture, the reversal of which would change the understanding of the events the speaker has related.

For example, I considered the following exchange between E9 and an Intermediate-High test taker as a request for narration. I underlined the examiner's speech to distinguish it from the test takers and I have removed names of people mentioned in the recording.

Ok so when you were in Amman, did any weird or funny story happen to you? Did anything weird happen to you?

Ah a weird story?

Like when you were in a taxi or on the street

Ah I was in the street ah my friend named _____ and I, we were walking in the street and we talked with Jordanians about anything and last week I was- we talked to one Jordanian ah about the month of Ramadan and ah my friend _____ he said to the Jordanian you know, you're poison and the Jordanian he's not- or he didn't understand and [my friend] he didn't know I didn't know the word "saam" ("sam" or poison) is different from fasting fasting in Ramadan "Saam"(=poison) is not- is a bad drink

Yes

And that was funny a little while later but at at the same time [to] the Jordanian it is not not funny*

Yes because he didn't understand what [your friend] wanted to say

After the examiner commented on the story, she then nominated another topic so I stopped transcription after the last line shown above.

In evaluating examiner questions and test taker responses for narration, I transcribed these from the moment the examiner nominated a topic to the narration request and response. Transcription was stopped when the examiner (or rarely the test taker) changed topics. In order to include the largest possible number of samples, a request for narration was defined as either a request for the test taker to tell a story from his or her personal experience or for a test taker to relate a story he or she had read. The total number of examiner requests for narration was larger as would be expected, but seven test takers declined to produce a response that could be categorized as narration or attempted narration and one test taker abandoned his narration attempt mid-response. After excluding declined requests and failed attempts, there were 19 requests left that fit my criteria and elicited an attempt at narration. This total also included one narration that a Superior rating level test taker and one that an Advanced-Low rating level test taker volunteered, in other words these narrations were not solicited by an examiner. Table 3.5 shows the number of narration requests in the data pool, according to rating level:

Table 3.5: Number of narration requests according to rating level

Rating	# of Narration Requests/total # of samples per rating
Superior	2 out of 8 (1 unsolicited narration)
Advanced-High	1 out of 11
Advanced-Mid	4 out of 17
Advanced-Low	4 out of 18 (1 unsolicited narration)
Intermediate-High	5 out of 18
Intermediate-Mid	4 out of 20
Intermediate-Low	0 out of 23
Total	20

It is unsurprising that there were no requests for narration at the Intermediate-Low rating level as these test takers generally struggled to understand and respond to simple questions. Test takers who received an Intermediate-Low rating level typically produced shorter responses than test takers at the higher rating levels. However, it was surprising to note that there was only one examiner request for narration in the Superior rating level recordings. As the number of Superior rating level interviews is small, it is unclear whether this can be generalized to other Superior rating level tests or whether this was merely a result of the sample size. Among the other rating levels, there were requests for narration in 9 of the Intermediate and 9 of the Advanced rating band tests, which presented the opportunity to contrast these sub-samples with one another with more confidence.

In order to offer observations about these narration attempts, I first examined the responses that seemed to be the most complete, in other words, the responses that appeared to satisfy or at least partially satisfy the narration requirements. I then re-read the transcripts and listened again to these test excerpts in each rating level band in an effort to determine what these responses had in common with one another. Once I finished looking for the commonalities among the most complete responses, I examined the responses that appeared to be the weakest. I defined the weakest as those that did not meet the minimum requirements of narration; the weakest were also often the excerpts with the least test taker speech and the most examiner speech. The results presented in Chapter 4 will be based on this process of comparison.

DESCRIPTION

In contrast to narration, the required elements of a description are more difficult to define. Benjamin Rifkin has even written that he thinks description emerges after narration in Russian L2 learners' speech and I find his anecdotal observation worthy of note (Rifkin, 2002, p. 466). I think description may pose a larger challenge than minimal narration because description may be used for various purposes in a non-testing context, such as offering an explanation or differentiating between objects or people. These purposes may or may not determine the components of a description.

Given that examiners could ask for a description related to any topic of their choosing, I was required to listen to all the samples in order to find the requests for descriptions across rating levels. I found that examiners asked for descriptions of hobbies, daily activities, friends and acquaintances, and cities or other places test takers had lived or visited. From among these topics, I chose to focus on requests for descriptions of cities. This was the most appropriate choice for three reasons: 1) these requests generally elicited more speech than requests for descriptions of other topics, 2) test takers mentioned themselves and their opinions less frequently than in other descriptions, and 3) the samples were similar in content, allowing for clearer comparisons between them. This also allowed for a consideration of what, if any, information and expressions formed characteristic answers for Advanced rating level test takers describing a city with which they were familiar.

In the OPI, a test taker attempts to provide a description in order to meet the requirements of the test so I could not form my working definition of description based on the communicative purposes for which the description was produced. Instead, my definition was necessarily very broad. My working definition was that Advanced-Mid and Advanced-High rating level test takers who were asked to describe a city would

produce some elaboration on that city using adjectives and phrases, and provide a description that differentiated that city from other well-known cities in the world. I expected that this elaboration might include comparisons between this city and other cities, but I did not expect to find this in all of the samples. I expected that some test takers might include personal anecdotes in their descriptions and that the set of adjectives produced would be limited to the most common like big, beautiful, crowded, or hot.

In order to separate requests for descriptions from other questions posed in the tests, I transcribed examiner questions that either included the word "describe" or were posed in a manner similar to, "How was [this city]? Tell me more about it." This kind of request was found in 34 different tests. Table 3.6 contains the number of description requests according to rating level.

Table 3.6: Number of description requests according to rating level

Rating	# of Description Requests/total # of samples per rating
Superior	2 out of 8
Advanced-High	2 out of 11
Advanced-Mid	4 out of 17
Advanced-Low	6 out of 18
Intermediate-High	7 out of 18
Intermediate-Mid	11 out of 20
Intermediate-Low	1 out of 23

Several test takers were asked to describe the same city. Table 3.7 shows the numbers of repeat requests for descriptions of the same city, categorized by the rating level at which each was found.

Table 3.7: Number of requests for descriptions of the same city

	Alexandria, Egypt	Cairo, Egypt	Damascus, Syria	Fez, Morocco
Superior	2			
Advanced-High	1	1		
Advanced-Mid			2	
Intermediate-High				2
Intermediate-Mid		4		

For example, I considered the following a request for a description. This excerpt is from a test in which an Intermediate-Low test taker mentioned having spent time in Marrakesh and the examiner, E9, asked for a description of that city. The examiner's speech is underlined to distinguish it from the test taker's.

> *Alright, you were in Marrakesh, I want you to describe to me Marrakesh, how was it?*
>
> *Sorry?*
>
> *Marrakesh is beautiful or … ?*
>
> *The weather is very hot.*
>
> *And what else? Tell me about the city.*
>
> *Yes, yes, in my opinion Marrakesh is a very beautiful city because you- because he- because I- because it you have- he has- it has the people are beautiful and nice and beautiful buildings but the weather is hot*

After this exchange, the examiner changed topics and I stopped transcribing this test.

QUALITATIVE METHODOLOGY FOR DESCRIPTION

In my analysis of the description sub-samples, I closely read and examined them in order to: 1) form an understanding of whether or not the content elements that Advanced rating level test takers produced in their descriptions differed from those produced by Intermediate and Superior rating level test takers and 2) to identify what I thought to be the most robust descriptions produced at each level, defined by the amount of intelligible production and examiner follow-ups. The first task was to understand whether there appeared to be a qualitative difference between the descriptions produced by Advanced rating level test takers and those produced by Intermediate and Superior rating level test takers. The second task was to identify the best achievements of both Intermediate rating and Advanced rating level responses, in order to provide an exemplar among these test takers' descriptions. However, I did not attempt to define an exemplary response at the Superior rating level because there were only two description samples at this rating level. Where possible, I also included a close analysis of city descriptions requested about the same city from different test takers.

In my initial review of the sub-samples, it appeared that examiner follow-up tended to include one or more of three question types; I present them here as discrete categories for the sake of clarity, but examiners sometimes used more than one question type in one turn. I took the example questions presented here from E2's exchange with an

Intermediate-Mid test taker. Examiner questions generally appeared to fall into the following types:

1. reiterations or rewordings of the same or a similar question on the same topic:
   For example, E2 opened one turn with: *And how was the city of Cairo, you know, how was it, what is a description of the city?* E2 then followed the test taker's response with a follow-up question that reiterated his first question:
   No, the city, how was it?
2. requests or reactions that suggested the examiner wanted clarification of what a test taker had just uttered:
   When E2 appeared not to understand, he said: *One more time please?*
3. questions that asked for more information on or an expansion of the same topic:
   And is the city of Cairo like the city of New York?

The first type of examiner follow-up was sometimes used to clarify a question to a test taker. However, if the first and second types of examiner follow-up were used together, then they appeared to be in response to an inappropriate test taker response or the use of a word or words that made the test taker's speech difficult to interpret. Based on these observations, I considered descriptions that included the first and second types of examiner follow-up—reiterations and requests for clarification—as weaker descriptions than those that were followed by requests for expansion.

Given that test taker abilities are evaluated using a holistic rating scale, I expected that the quality of the individual descriptions would vary, even within rating levels. I therefore present some discussion of the weakest descriptions—as defined by length, and

my observations of intelligibility and examiner response—in Chapter 4[17]. The differences were particularly pronounced among Intermediate level test takers, but were also noteworthy at the Advanced level and are therefore discussed in the next chapter.

[17] See page 97 for Advanced level test takers or page 126 for failed narrations at the Intermediate rating levels.

CHAPTER 4: RESULTS AND ANALYSIS

In the previous chapter, I described the methods used to generate my results. I interpreted lexical breadth as a measure of the number of words and average words per minute produced by test takers at different rating levels, and I explained the calculation of TTR for lexical richness. Lexical frequency was measured using Buckwalter and Parkinson's Arabic frequency dictionary. I also explained how I defined description and narration and how I separated requests for description and narration from other examiner questions in the test.

In this chapter, I will discuss the results and analysis of my data, beginning with the quantitative descriptors. First, I will present the range of words and number of words produced per minute, separated by rating level for the 28 full-length samples. Next, I will present the number of tokens and TTR measurements from the full-length samples and sub-samples of description and narration, discussing the differences between the Advanced rating levels and the Intermediate and Superior rating levels for both measures. I will also discuss word frequency data for words that test takers produced in common among description and full-length samples from the same rating level. Lastly, I will present my observations from the description and narration qualitative analyses.

QUANTITATIVE DESCRIPTORS: WORD PRODUCTION RANGE AND WORDS PER MINUTE

The quantitative descriptors were gathered to answer my first research question on lexical breadth:

1. What are the average words and words per minute produced by Advanced-Mid rating level test takers in a subset of the OPIs under consideration? Do Intermediate-Mid rating level test takers produce fewer words and words per minute than Advanced-Mid rating level test takers? Do

Superior-level test takers produce more words and more words per minute than Advanced-Mid speakers?

I produced ranges for all eight Superior level tests and for ten randomly selected tests from the Advanced-Mid and Intermediate-Mid levels. I reviewed all transcripts for accuracy and then counted the number of words produced by each test taker. In my review of the Superior level test taker production, I found that one test taker appeared to produce far more words than the other seven, even though his recording was not the longest. Upon reading this test taker's transcript again, I found that this speaker used the word *ya^{C}nii* (a filler that is similar to "like" or "you know") very frequently; a subsequent count showed that he produced it 296 times over the course of his test. I then reviewed the other transcripts and counted the number of times the other test takers produced this filler. Although none of the other speakers produced it as frequently as the first Superior level speaker I examined, I adjusted all of the word ranges to exclude this speech filler to ensure the comparability of the numbers. Table 4.1 contains the raw word ranges and an adjusted word range with *ya^{C}nii* removed:

Table 4.1: Raw and adjusted word production ranges in full-length tests according to rating level

Level	Word Production Range	Adjusted Word Production Range
Superior	1192-2016	1185-1720
Advanced-Mid	913-1942	910-1859
Intermediate-Mid	604-1051	604-1051[18]

These raw and adjusted word production ranges indicate that there is a clear difference between Intermediate-Mid and Advanced-Mid word production ranges. However, there is no discernible difference between the word production ranges of the Advanced-Mid speakers and the word production ranges of the Superior rating level test

[18] Three of the Intermediate-Mid test takers used *ya^{C}nii* but they were not among the lowest or highest word producers so the adjustment to the word count is not apparent in these word production ranges.

takers. The Advanced-Mid test takers' word production range showed a larger spread than the Superior test takers, with Advanced-Mid test takers producing a range that encompassed the narrower range of Superior test takers.

Because the interviews varied in length and were conducted by different examiners, I also calculated the average number of words per minute produced by each test taker. Table 4.2 shows the average words per minute (WPM) (with the filler *ya^{C}nii* removed), which illustrates more clearly the difference between the Intermediate-Mid and Advanced-Mid test takers' lexical breadth.

Table 4.2: Average words per minute and range of average words per minute produced per rating level

Level	Range of Average Words Per Minute	Average Words Per Minute Per Rating Level
Superior	45.45 – 68.01	55.76
Advanced-Mid	34.98 – 71.03	52.88
Intermediate-Mid	29.51 – 40.19	34.85

The difference between the WPM produced by Superior rating level test takers and those produced by Advanced-Mid test takers still appears to be minimal. The spread was larger for Advanced-Mid word production averages than for the Superior level, and this could have made the average appear higher. To address this, I also calculated the median words per minute for each level, shown in Table 4.3 below.

Table 4.3: Median words per minute per rating level

Level	Median Words per Minute
Superior	51.43
Advanced-Mid	55.21
Intermediate-Mid	35.22

The medians indicate that Advanced-Mid speakers typically produced more words per minute than Superior rating level speakers, not less, as might be expected based on Malone's and Read and Nation's separate findings[19]. It appears that the number of words does not automatically rise when a speaker is rated at the higher level of Superior. This seems to indicate that reaching the Superior rating level is not just an issue of the quantity of production.

In order to test whether this difference was statistically significant, I conducted independent-samples t-tests to compare the Advanced-Mid rating level's individual WPM to the WPM of the Intermediate-Mid and Superior rating level test takers. There was no significant difference between the Advanced-Mid WPM ($M = 46.87$, $SD = 10.62$) and Superior rating levels' WPM ($M = 47.44$, $SD = 12.97$); $t(16) = -0.10$, $p = 0.91$. This makes it appear that WPM is not a factor that distinguishes the Advanced-Mid rating level speakers from the Superior rating level speakers in this test. In contrast, I found a significant difference between the Advanced-Mid rating level's ($M = 46.87$, $SD = 10.62$) and the Intermediate-Mid rating level's WPM ($M = 28.97$, $SD = 4.24$); $t(11.79) = 4.95$, $p < .001$. This suggests that WPM is a factor that can distinguish between Advanced-Mid rating level and Intermediate-Mid rating level speakers.

[19] It should be noted that both these researchers used more controlled data in their research into the simulated OPI (SOPI) and the IELTS tests as both the SOPI and the IETLS tests appear to be given in more rigid ways than the typical ACTFL OPI.

QUANTITATIVE DESCRIPTORS: TTR IN FULL-LENGTH SAMPLES

I used WordSmith Tools to produce the TTR for each of the full-length transcripts that exceeded 500 tokens. I chose 500 tokens in order to include the largest possible data set[20]. I generated these measures to answer my second research question:

2. What is the lexical variation in the Advanced-Mid samples as measured by type-token ratio (TTR)? Is this diversity higher or lower than the lexical diversity of test taker samples at the Intermediate-Mid and Superior rating levels?

Table 4.4 shows the range of TTRs for each test taker rating level.

Table 4.4: Range of TTRs for test rating levels

Superior	45.4-56.4
Advanced-Mid	40.8-53.2
Intermediate-Mid	37.2-47.2

I ran independent-samples t-tests on the TTRs for each rating group. The results were similar to the WPM measure. The difference between the Advanced-Mid TTR (M=45.14, SD=4.19) and the Superior TTR (M=49.5, SD=4.39) was not significant $t(16)=1.57$, $p=.137$. However, the difference between the Advanced TTR (M=45.14, SD=4.19) and the Intermediate-Mid TTR (M=41.34, SD=3.54) was significant $t(18)=2.19$, $p=.042$. This suggests that Advanced-Mid test takers are not only producing more words than Intermediate-Mid test takers but also producing more varied words.

As stated earlier, there were 33 description sub-samples and 20 narration sub-samples. The total number of tokens for these sub-samples varied widely. The description

[20] This threshold of 500 tokens excluded one test at the Intermediate-Mid level. Therefore, I tested the TTRs of 8 Superior rating level tests, 10 Advanced-Mid rating level tests, and 9 Intermediate-Mid rating level tests.

sub-samples were all 300 tokens or less; the narration sub-samples were also under 300 tokens except for two outliers. One Advanced-High test taker produced 517 tokens and one Intermediate-Mid test taker produced 357 tokens. Table 4.5 shows the range in the number of tokens per rating level.

Table 4.5: Range of Tokens in Description and Narration Sub-Samples

	Description	Narration
Superior	138-149	236-284
Advanced-High	151-230	517
Advanced-Mid	110-222	73-161
Advanced-Low	97-299	116-222
Intermediate-High	86-193	41-139
Intermediate-Mid	26-194	98-357
Intermediate-Low	29	No samples at this level

I also calculated a TTR for the first 100 tokens for every sub-sample that included 100 tokens or more; I chose 100 tokens in order to include as many of the sub-samples in the data set as possible. In the descriptions, 21 sub-samples contained 100 tokens or more. In narration, 16 of the 20 sub-samples contained 100 tokens or more. The TTRs for these short samples did not appear to follow any pattern. Table 4.6 shows the range of TTRs for the sub-samples that exceeded 100 tokens.

Table 4.6: Range of TTRs for description and narration sub-samples of 100 tokens or more

	Description	Narration
Superior	49-66	63-65
Advanced-High	63 (samples had the same TTR)	64 (1 sample only)
Advanced-Mid	53-59	64-77
Advanced-Low	58-65	58-67
Intermediate-High	53-61	53-55
Intermediate-Mid	58-70	48-63
Intermediate-Low	Sample did not exceed 100 tokens	No narration samples

There were too few sub-samples to run independent t-tests, and the number of tokens and TTRs did not suggest a pattern worth testing, even if there had been a sufficient number.

QUANTITATIVE DESCRIPTORS: SHARED VOCABULARY ACROSS CITY DESCRIPTIONS

In regard to vocabulary shared among test takers, my initial working assumption was that test takers asked to describe the same topic would produce an appreciable number of the same words, even when describing different cities. In order to test this hypothesis, I put the description transcripts into WordSmith Tools to generate a list of the words test takers used most commonly in their descriptions, grouped by rating level. The Advanced-Low, Advanced-Mid, and Advanced-High rating level test takers produced 12 sub-samples in total. Using these, I generated a list of the words that Advanced rating

level test takers used in approximately half or more of their descriptions[21]. I chose the cut off of half of all sub-samples because this would produce a shared vocabulary pool from six or more test takers in the Advanced rating levels and nine or more test takers in the Intermediate rating levels; given the small number of sub-samples I chose half in order to gather a larger pool of shared vocabulary than would be possible if I only examined vocabulary that was produced in all the description sub-samples[22]. This resulted in 28 words that six or more Advanced rating level test takers produced in common[23]; the majority of these were function words. There were only six words with semantic content produced across half or more of these descriptions. These words were: "city" (both definite and indefinite versions), "beautiful" (both masculine and feminine singular versions), "a lot" (both masculine and feminine singular versions), "the people," "according to" (as in "according to me"), and "thing."

The Intermediate rating level test takers produced 19 descriptions across the Intermediate-High, Intermediate-Mid, and Intermediate-Low groups. They produced only 11 of the same words in half or more of their descriptions[24]; the majority of these words were also function words. Intermediate rating level test takers produced only two shared words with semantic content across their descriptions, which were "city" (both definite and indefinite versions) and "the people." The Intermediate rating level produced all of

[21] The Intermediate rating level samples were 19 total so I used 9 as the cut off point for the list of shared vocabulary words.

[22] If I had chosen to consider only the words produced by all 12 test takers in the description sub-samples from the Advanced rating levels or all 19 of the Intermediate rating levels, this would have only resulted in the word "and."

[23] The numbers in this section are the totals of words added in Arabic but, of course, the English equivalents may require more than one word to translate. See Appendix A for the complete list of shared words produced across the Advanced rating level descriptions.

[24] See Appendix B for the complete list of words produced in Intermediate rating level descriptions.

the same shared words as the Advanced rating level test takers, except for the addition of the question word "how."

At the Superior level, there were only two description samples and both test takers were asked to describe Alexandria, Egypt. I generated a list of words produced by both test takers in order to understand which words these two test takers used in common; the list of shared vocabulary extended to 18 tokens. Nine of these words were the same as the ones produced by the Advanced rating level speakers. The Superior rating level speakers produced only four shared words with semantic content that were not produced by the Advanced rating level speakers. These were: "other" (feminine singular version), "Alexandria," "the sea," and the definite adjectival form of "Egyptian."

The comparisons of the words produced in the description sub-samples seem to indicate that my assumption of a discernible shared descriptive vocabulary pool was not realistic, at least among samples grouped together at the Advanced and Intermediate rating levels. In order to test this further, I used WordSmith Tools to generate lists of words used by test takers to describe the same city. This allowed me to remove different cities as a factor that might contribute to the variety in vocabulary being produced by test takers. To this end, I generated combined vocabulary lists for three sets of descriptions. The first was from two Advanced-Mid rating level test takers who were asked to describe Damascus, Syria. The second was from the two Superior rating level speakers and one Advanced-High rating level speaker who were asked to describe Alexandria, Egypt. The third list was generated as a comparison to the higher levels, using two Intermediate-High test takers' descriptions of Fez, Morocco.

I then compared the shared vocabulary that was produced by test takers describing the same city[25] with the shared word lists from test takers describing different cities. There was very little difference. The test takers describing the same city used a few words in common that the test takers describing different cities did not. The Advanced-Mid test takers contributed only four new shared content words: "Syria," "Syrian," "Damascus," and "culture." The combined Superior and Advanced-High list was the same as the list produced when I compared the shared words of the Superior test takers. The Intermediate-High shared word list included the most new words; however, this group was still very small. In addition to "Fez" and "Moroccan," the two Intermediate-High test takers also produced four other words in their descriptions: "history," "old" (both masculine and feminine singular versions), "place," and "there is/there exists." This additional analysis lent support to the impression that these test takers did not use a shared pool of descriptive vocabulary in these test recordings.

In order to widen my search for shared vocabulary, I generated combined word lists for the complete test transcripts from Advanced-Mid, Intermediate-Mid, and Superior. Since I had ten samples from Advanced-Mid and Intermediate-Mid, I examined the words that were used in eight or more of the samples. In the Superior group, I had 8 samples, so I examined words that were used in six or more of these samples. I chose both of these numbers so that the shared words would have been produced in 75-80% of the samples under consideration. The results were very similar, i.e. function words made up the majority of shared vocabulary and these words were largely the same between the three groups. The only difference I found between the three was that Intermediate-Mid produced the smallest shared pool of 32 words in 8 of 10 transcripts, the Advanced-Mid

[25] For a complete list of the shared vocabulary among test takers describing the same city, please see Appendix C.

shared 57 words in 8 of 10 transcripts, and the Superior rating group producing the most shared vocabulary with 74 words produced in six of eight transcripts.

I also expanded this to include shared words that were produced in 50% of the transcripts in order to see if this affected which rating level produced the most shared words. It did not. The Superior rating test takers still produced the most words in common with 194 words produced in four of eight transcripts. The Advanced-Mid test takers' produced fewer with 160 shared words and the Intermediate-Mid test takers' produced the least with 112 shared words respectively in 5 of 10 transcripts.

QUANTITATIVE DESCRIPTORS: FREQUENCY RANKINGS OF SHARED WORDS

Finally, I examined the shared words' frequency rankings, using Buckwalter and Parkinson's Arabic frequency dictionary (Buckwalter & Parkinson, 2011), in order to answer my third research question:

> 3. How many shared words produced by learners at the Advanced rating levels are from beyond the 2,000 most frequently used words in Arabic according to Buckwalter and Parkinson's frequency dictionary (Buckwalter and Parkinson 2011)?

Buckwalter and Parkinson's dictionary contains the 5,000 most commonly used Arabic lemmas culled from a corpora of 30 million words, of which 10% were spoken data; a higher ranking indicates a less commonly used word. My assumption was that less commonly used words would correlate with a higher rating level in this speaking test. However, I did not find a connection between Buckwalter and Parkinson's frequency ranking of the shared vocabulary in this data and test takers' rating levels. In the combined word list for the Advanced rating level descriptions, 27 of the 28 shared words were from among the 500 most common words according to Buckwalter and Parkinson.

The only exception was "like" or "you know" which was ranked 751 and was used overwhelmingly as a filler in test taker speech. The Intermediate and Superior shared vocabulary lists for city descriptions also contained frequently used words, the vast majority of which were from the 500 most commonly used words. I also examined the shared vocabulary lists from the complete test transcripts. Similarly, most words were from among the 500 most commonly used. The words that were not—like "Arabic" and "I study" which are ranked 2509 and 1066 respectively—could arguably be considered high-frequency for L2 speakers of Arabic who have learned the language in an instructional context.

In summary, the findings indicated that the majority of the shared words produced by Intermediate, Advanced, and Superior rating speakers in descriptions and selected full-length tests were from the 1,000 most commonly used words according to Buckwalter and Parkinson's frequency dictionary. While there were some exceptions among the words produced, these exceptions were words that Arabic L2 learners would be likely to have learned early in the course of their Arabic studies or were often proper nouns like the names of cities or countries. In other words, I did not find support for my assumption that lower frequency words would be found in the speech production of test takers who received higher ratings.

QUALITATIVE OBSERVATIONS

The following section answers my fourth and final research question:

> 4. What qualitative observations can be made about test takers' narration and description attempts at the Advanced rating levels? How do these attempts compare to narration and description by test-takers at the Intermediate and Superior rating levels respectively?

I will begin with the description samples and then discuss my observations of the narration samples. Since my focus is on the Advanced rating levels, I will present observations of Advanced rating level test takers first and then discuss Intermediate and Superior rating level sub-samples. As there was only a general definition of "description" from which to begin, my aim in analyzing these sub-samples was to see if they were similar to and different from one another in systematic ways. My working assumptions were that city descriptions provided by Advanced rating level test takers would provide some elaboration in the form of adjectives and phrases like "in general" or "in my opinion," and that the descriptions would distinguish the city being described from other similar cities.

However, I also began my analysis with the assumption that the sub-samples could be largely similar among rating levels, showing little that distinguished one rating level response from another. This was an important assumption to include so that I would not succumb to confirmation bias, i.e. allowing my expectations to influence what I found in the data. In addition, the OPI is a holistic measure and test taker performance can be expected to fluctuate naturally across different tasks. Given these two facts, sub-samples taken out of context of the entire testing performance might not be significantly different from one another and might not reflect the best performance the test taker was capable of.

Before turning to the observations, it should be noted that Intermediate rating level speakers were not always able to respond to the request for description in an intelligible manner. Of the Intermediate rating level speakers who were able, most tended to insert first-person language frequently (for example, often mentioning whether they liked or disliked a city), especially at the beginning of their attempted descriptions, and to start and stop their speech frequently as they searched for words or word forms. Their descriptions typically focused on basic information about the city like whether it was big

or small, whether it was crowded or not, and what the buildings and streets looked like. Their attempts typically did not produce language that could serve to distinguish one city from another, other than by general categories like size. Intermediate level speakers were also likely to give indications that they were having difficulties retrieving or producing key words, making basic information in their descriptions harder to understand than those at the higher rating levels.

In contrast, the Superior level speakers presented the same content as the Intermediate and Advanced level speakers, but also included geographic location or limited historical references. In addition, one of the Superior speakers acknowledged his listener's potential reference points in his response, and the other Superior speaker mentioned the dialect and languages of the inhabitants of the city. Speakers at lower rating levels did not include this kind of information in their description responses.

My main findings at the Advanced rating levels were that Advanced city descriptions differed most from Intermediate and Superior rating level descriptions in terms of lexical use and resulting content. Advanced rating level speakers tended to include content similar to the Intermediate level speakers, but they typically added more information and responded in clearer ways. They also produced descriptions that were less similar to one another, and less similar to those produced at the Intermediate level. Some Advanced-Low speakers still mentioned their own experiences in the city, but Advanced-Mid and Advanced-High speakers tended to present their descriptions using less first-person language. In the following section, I will present some examples from each rating level and explain in more detail my judgment of that sample.

Advanced Descriptions

There were 12 sub-samples from the data for the Advanced rating level group: 6 in Advanced-Low, 4 in Advanced-Mid, and 2 in Advanced-High. In contrast to the Intermediate speakers, the Advanced-level test takers produced different content and their attempted descriptions can be more easily distinguished from one another, an indication of their more varied vocabulary use when compared with the Intermediate level. This may be part of the reason why the combined word lists did not reveal much in regards to a shared vocabulary pool. Compared with the Intermediate samples, the Advanced samples seemed less generic in nature. It appears that Advanced test takers are at an ability level that allows them to produce what can be considered consistently more successful—albeit limited—descriptions.

As noted above, the Advanced test takers included more of their personal perspective at the "low" sub-level than at the "mid" or "high" sub-levels. In Advanced-Low, 5 of the 6 test takers used first person language, focusing on what they did or not did not like about a particular city and the activities they did there. In contrast, only one Advanced-Mid and one Advanced-High test taker included first-person language. Moreover, the Advanced-High speaker included only the comment, "I prefer this city," and then focused the remainder of the description on the city itself.

In the section that follows, I will present three of the strongest examples of speakers at the different sub-levels, in order to support the characterizations of the Advanced rating level I have presented above. I will begin with the strongest Advanced-Low rating level and proceed through the sub-levels. In the transcripts, examiner contributions are underlined to differentiate them from test taker production, and an

English translation of the Arabic is included. I attempted to include in the English translations the same repetitions, similar lexical errors, and abandoned clauses as those found in the original Arabic speech. I also attempted to preserve some dialectical pronunciation in the transcripts when this pronunciation was present in the test taker's speech. Errors are marked with asterisks in the original and in the English translations. Empty underlined spaces represent personal or place names that have been redacted from the transcripts.

The first example is from an Advanced-Low test taker and was elicited by E1 at about 12:40 in the interview. Bold indicates words that were said in English:

ربما ربما سنعود إلى هذا الموضوع بعد قليل اللغة العربية الاهتمام بها لكن دعينا نتكلم عن مدينة بيركلي هل يمكن أن تصفي لي تلك المدينة الجميلة؟

طبعا أنا دلوقتي الآن أنا عشت في بيركلي خمس سنوات تقريبا وأنا والله احب بيركلي كثيرا أنا من نيويورك وفي البداية ما كنت اريد ان انطلق إلى انتقل إلى مدينة خارج نيويورك ولكن عندما وصلت إلى بيركلي شعرت بالراحة في هذه المدينة الجميلة يعني هي مدينة صغيرة إلى حد ما بس فيها كل الأشياء المهمة لها المدن الكبيرة مثل نيويورك وسان فرانسيسكو بجانب مدينة بيركلي على طول فيعني بالنسبة للأكل فيه أكل هناك أكل ممتاز من من كل بلد ويعني بالنسبة للحياة الثقافية يعني في كل شيء موجود أو في بيركلي أو في سان فرانسيسكو وهي مدينة قريبة جدا بالنسبة للمواصلات يعني سهل سهل جدا السفر في في هذه المدينة وأنا من نيويورك ولا أعرف كيف نقول اقود سيارة فالحمد لله ليس هناك احتاج أو حاجة للسيارات هناك أنا عندي عجلة وخلاص وايه تاني الحياة يعني الحياة يعني صحية جدا كل الأكل الخضار والفاكهة الفواكه جميل جدا ويعني أحسن من نيويورك لأن سرع الحياة أبطٌ بكثير

ماذا؟

بطيء يعني أبطا أبطا من من نيويورك

وهل هذا شيء جيد؟

بالنسبة لي آه يعني لأنني كنت يعني كبرت في هذه المدينة وما عرفت ما عرفت أن السرعة سرعة

نيويورك كانت مجنونة ولكن عندما انتقلت إلى مدينة أخرى وكان عندي هذه التجربة أن أعيش في مدينة شوية أكتر يعني بطيء فبطيئة فأنا أفضل هذا

قلتِ هناك توافر كبير في الخضروات والفكهة كيف هي الأسواق في بيركلي

بمعنى الأسواق يعني

الأسواق في الهواء الطلق هي أسواق في مناطق مغلقة بنايات مثل يعني أسواق الخضروات العادية أم هي أسواق مفتوحة في الشارع؟

هناك هناك أسواق هناك ثلاث أيام في الاسبوع هناك أسواق مفتوحة للهواء بس و- ولكن أنا كثيرا أذهب إلى سوبرماكت واحد هو والله مشهور للخضروات والفواكه لأن اسم السوبرماركت البيركليبول وفيها كل شيء والله ويعني مثلا هناك اثناشر نوع من المانجة مثلا من من ثمانية يعني بلدان مختلفة وفأنا دائما وأنا أسكن قريبا جدا من هذا السوبرماركت

انت محظوظة جدا يا _____

I know
والله أنا كل يوم حمدلله أشكر الله يعني

هذا شيء جميل جدا تمام

Let's talk about the city of Berkeley. Can you describe this beautiful city for me?

Of course, I now now[26] *I have lived in Berkeley for five years now approximately and I really like Berkeley a lot I am from New York and in the beginning I didn't want to take off* to move to a city outside New York but when I arrived in Berkeley I felt comfortable in this beautiful city like it is a small city to a certain extent but it has a lot of important things like in big cities like New York and San Francisco is right next to the city of Berkeley so in terms of food there's there is great food from every country and in terms of cultural life you know everything is here either in Berkeley or in San Francisco and it is very close in terms of transportation you know it's very easy to travel in this city*

[26] This speaker uses the dialect version of "now" and then uses the formal version of "now" here.

and I am from New York and I don't know how to- how do we say to drive a car so thank God there is no need[27] *or need for cars there I have a bike and that's it and what else? Life is you know life is you know very healthy- all the food the vegetables the fruit the fruit*[28] *is very nice and you know it's nicer than New York because the speet* of life is a lot *slow-*

What?

Slow- you know slower slower than New York

And is that a good thing?

In my opinion, yeah, because I- you know grew up in this city and didn't know didn't know the speed of New York was crazy but when I moved to another city and I had this experience of living in a city that was a little you know slower so I prefer that

You said there are a lot of fruits and vegetables available. How are the markets in Berkeley?

The markets meaning …?

The open-air markets- the markets are in an enclosed building like normal vegetable markets or are the markets open on the street?

There are there are markets there are three days a week there are markets open to the air but and but often I go to a supermarket and really it's famous for vegetables and fruits because the name of it is Berkeley Bowl and it has everything really and you know like there are twelve kinds of mango for example from from eight like different countries and so I always- and I live very close to that supermarket*

---

[27] This is one form of the word "need." The test taker supplies another, more common form immediately following.

[28] This is another instance in which the test taker uses a word favored more in dialect before producing a more formal equivalent. Both words are pronounced correctly and are appropriate to the context; they only differ in terms of register.

You are very lucky!

I know

I swear every day thank God- I thank God!

That is very nice, ok

This test taker's initial description lasts approximately two minutes. In it, she focuses on her personal experience, but also provides more information than test takers at the lower levels. She includes: 1) her opinion, 2) how the city compares to her hometown, 3) some information on its proximity to San Francisco, 4) that the transportation is good enough that she does not need to drive, and 5) that the lifestyle is healthy because there are lots of fruits and vegetables available. The follow up questions from the examiner appear to be aimed at expanding the information the test taker has given. This also seems to imply that her description was understandable and provided sufficiently clear information, allowing the examiner to focus on expansion questions rather than clarification questions.

At the Advanced-Mid level, one of the strongest descriptions was also elicited by E1. It occurred about twelve minutes into the interview.

طيب أنتِ كنت في المغرب ما هي في رأيك أجمل مدينة مغربية؟

هذا سؤال سهل لأنّ أجمل مدينة هي تارودانت تارودانت

تارودانت نعم

هل يمكنك أن تصفي المدينة بالتفاصيل؟

نعم تارودانت يعني اسمها جدة مراكش لأنّ* أقدم من مدينة مراكش ومراكش مدينة مشهورة جداً

وفيها تنمية* وسياسية أو السياحة كثيرة الآن ولكن تارودانت أقدم وأجمل من مراكش ولكن ما زال يعني صغير شوية بالنسبة للسياحة وهذا جيدة ليس فيها ازدحام والثمن* غالية جداً وذاك الشي لأنّ هيك شيء يعني ليس فيها سياسة-* سياحة كثيراً ولكن فيها يعني أسواق ممتازة فيها يعني كل شيء من صناعة تقليدية

like

منتوج صناعة تقليدية يعني وأيضاً ماكلة ممتاز وفي وسط مدينة في وسط منطقة امازيغية ولكن المدينة عندها تاريخ طويل من يعني حضرة العرب ولذلك هناك يعني العرب والامازيغية والمأكلة والثقافة في هذه المدينة يعني جميل جداً

وكيف هي الشوارع والبيوت مثل مراكش أو مختلفة؟

مثل مراكش ولكن في مراكش هناك بيوت قديمة جداً ولأنّ الناس الـ .. خارجية في مراكش هناك يعني تحفيظ كثير من هذه البيوت وممكن

مرّة ثانية؟ هناك؟

هناك يعني الناس الذين يحافظوا البيوت في نفس الشكل ككانت البيوت في الماضي ولكن بالنسبة للناس في تارودانت كان بيوت كثير من الناس في تارودانت يريدون أن يعني يغيّروا بيوتهم إلى بشكل معاصرة يعني مع يعني كونكريت وهذا شيء والآن المدينة تغيّر في الماضي كان حدائق عامة في كل بلاصة وشجرة البرتقالي والرومان وكل شيء في كل ميدان ولكن أقل وأقل الآن أعتقد

أنا إن شاء الله سأزور هذه المدينة إذا ذهبت إلى المغرب يعني مدينة جميلة
نعم أحسن مدينة

Well, you were in Morocco. What in your opinion is the most beautiful Moroccan city?

That is an easy question because the most beautiful city is Taroudant Taroudant yes

Can you describe the city in detail?

Yes, Taroudant you know its name is "grandmother of Marrakesh" because older than the city of Marrakesh and Marrakesh is a very famous city and it has growth and politics and much tourism now but Taroudant is older and more beautiful than Marrakesh but now it is still a little small in terms of tourism and this is good there isn't crowdedness and high prices* and that sort of thing because of that you know it doesn't have politics-* tourism a lot but it does have like great markets in it you know everything of traditional manufacturing* ***like*** *production traditional production for- you know you know and and also great food and and in the middle of a city- the city is in the middle of an Amazigh area but the city has a long history of from the time of Arab settlement and therefore there are the Arabs and the Amazigh and the food and the culture in this* city this city are very beautiful*

And how are the streets and cities- like Marrakesh or different?

*Like Marrakesh but in Marrakesh there are very old houses and because of the outside people in Marrakesh there is * a lot of memorization* from a lot of houses and maybe*

One more time? There is…?

*There is umm you know the people who preserve the houses in the same appearance as the houses were in the past but in regards * the people in Taroudant *it was a lot of houses from people in Taroudant who want to you know change their houses into by*a modern style like with the concrete and that thing and now the city change* in the past, there were public parks in each square and orange trees and pomegranate and every thing and in each square and but less and less now I think*

Hopefully I'll visit this city if I go to Morocco you know- a beautiful city!

Yes, yes, the best city

This is one of the strongest examples of Advanced-Mid descriptions because the test taker provided more information and more varied content about the city than other speakers at the Advanced and Intermediate rating levels in less than 3 minutes and 30 seconds. She tells the examiner: 1) Taroudant is the most beautiful city in Morocco in her opinion, 2) it is known as the "grandmother of Marrakesh," 3) it is smaller than Marrakesh, 4) it is known for traditional crafts and markets, 5) it has a long history from the arrival of the Arabs, and 6) that both Arab and Amazigh cultures are found there. She is also able, in a limited way, to respond to the requests for expansion by discussing the houses and attempts to preserve them from current inhabitants' efforts to renovate or otherwise change them.

I will now turn to one of the strongest examples of a city description from an Advanced-High test taker. This description was elicited by E2 approximately 11 minutes into the interview. Words in bold were spoken in English in the recording:

فقلتَ إنكَ أيضاً يعني كنتَ في دمشق ثم ذهبتَ إلى القاهرة واكملتَ الدراسة في القاهرة نعم؟

نعم

حدثني عن مدينة القاهرة كي انا ما زرتُها ما زرتُها لا أعرفها فأريدُ وصفاً يعني مفصلاً عن مدينة القاهرة

مدينة القاهرة هي مدينة جميلة جداً هي مدينة كبيرة واسعة فيه ناس كثير في زحمة في كل حتة فعدد السكان تقريباً خمسة وعشرين مليون وفيها بالنسبة لي هي متحف ف فيه آثار رومانية يونانية فرعونية إسلامية كل هذه الآثار موجودة في نفس المدينة فممكن أنت تمشي في المدينة وفيه بنايات مثلاً إسلامية جنب آثار رومانية وآثار يونانية في نفس الشارع مع بعض وفيها مثلاً بالنسبة للأكل في مطاعم مختلفة الأكل ممتاز أكل لذيذ هناك وهي رخيصة بالإضافة إلى ذلك بتتميّز بتتميّز القاهرة ب

coffee shops

قواهي شعبية فيه قواهي شعبية والناس قاعدين وبتشرب أرجيلة وبتتكلم فالجو حلو كثير يعني هي مدينة حلوة كثير وفيه جامعات فيه مكتبات في حاجات مختلفة ممكن أنت تمشي ممكن أنت تزور منطقة كل يوم لمدة سنة وأنت ما تخلصش

لسه فيه حاجات مختلفة وجديدة كل يوم

ليس فيه أو فيه؟

لسه يعني فيه فيه

So you said, you know, you were also in Damascus and then you went to Cairo and you finished your studies in Cairo, yes?

Yes

Uh, tell me about the city of Cairo as if I had never visited it- I've never visited it and I don't know it so I want you know a detailed description of the city of Cairo

Uh the city of Cairo is a very beautiful city it is a big- wide city there are a lot of the people it's crowded everywhere since the population is approximately 25 million and in it in my opinion it is a museum since there are Roman, Greek, Pharaonic, Islamic ruins all these ruins are present in the same city so it's possible for you to walk in the city and there are Islamic buildings for example next to Roman ruins and Greek ruins in the same street together in it for example in terms of the food in different restaurants the food is delicious food there it is cheap in addition to that Cairo is distinguished by ***coffee shops*** *the popular ceffoo* houses there are popular ceffoo* houses and the people are sitting and you smoke argeelah and you're talking so the weather is very nice and there are universities and there are libraries and there are different things maybe you walk maybe you visit a neighborhood every day for a year and you haven't finished ahh there are still new and different things every day*

There aren't or there are?

There still you know there are there are

This test taker produced common features of the content of both an Intermediate rating level description and an Advanced level description. He included the beauty, size, and weather of Cairo as Intermediate level test takers commonly did, but he also referred to the history and ruins found there. He listed good food and varied restaurants as attractions, and his description moves from a list of the attributes to include a description of what the listener might enjoy, something that was not found in the majority of the lower-level descriptions.

The Advanced rating level speakers appeared to include more information in their descriptions than the test takers at the Intermediate rating level, as I will explain below. Advanced rating level speakers' inclusion of more varied information also made their descriptions more easily distinguishable from one another. In this respect, these descriptions were also more successful than those found at the Intermediate rating levels. However, they were still limited in the scope of information found in them – generally limited to comments on size, appearance, weather, and history – and occasionally contained errors in pronunciation or vocabulary selection that prevented these descriptions from being as easily understood as descriptions elicited from test takers who were rated at the Superior level.

Intermediate Descriptions

The lowest level Intermediate description attempt was elicited by E9 at approximately 11:30 in an interview with a test taker who was subsequently rated

Intermediate-Low. The test taker mentioned that he had visited Marrakesh and the examiner asked him to describe the city. This test taker produced an attempt at description that included his personal opinion of the city ("for me, Marrakesh is beautiful"), a simple statement about the people and buildings of the city ("good people and beautiful buildings"), and a repetition of a comment about the weather ("but the weather is hot"). The entire exchange – including the examiner's prompting – takes approximately one minute. The test taker's description is 29 tokens long (including searching for appropriate word forms), making it one of the shortest attempts found in the data set. He does not include any information that differentiates this city from other cities in the world, other than the fact that the speaker considered it beautiful.

The 11 Intermediate-Mid level speakers tended to concentrate on the appearance of the city (with 6 commenting on its beauty or size) and their personal experience of the city (whether they liked it or not, including two impressions of the city's inhabitants). Two also mentioned weather, a topic that was found in only a few samples across the Intermediate and Advanced levels. Exceptions to this overall characterization of Intermediate level description were found in three responses, in which test takers volunteered more information or evaluation. The longest exception came from a test taker who volunteered that he goes to the city he is attempting to describe in order to flee his university, which is like a prison. His description extended to 93 tokens. In another test, the test taker included geographic information, without being asked to do so by the examiner, and a third test taker tells the examiner not to go to Morocco.

Additional information was included in test takers' descriptions if and when examiners asked more specific questions, either about the geographic location, economy, or reasons why a city was particularly well-known. I consider these responses to show evidence of test takers' abilities to answer specific questions about the cities at the

Intermediate level. However, I excluded these from my characterization of city descriptions at this level as the test takers did not volunteer this information in their initial responses, but rather these pieces of information were elicited from them by examiners' follow-up questions.

Responses to the requests for description at the Intermediate-Mid level can best be described as largely unsuccessful. First, the responses were typically difficult to interpret because they often included either multiple lexical errors or lexical errors at key junctures in the communication that made understanding the test taker responses extremely difficult. Second, the responses largely consisted of a string of place names or disconnected statements that required repeated examiner prompting to elicit. This appears to change for some test takers at the Intermediate-High level as I will explain below.

Intermediate-High Descriptions

There were 7 sub-samples of Intermediate-High descriptions. Intermediate-High level speakers appeared to encounter difficulty in communicating meaning in their description attempts, and did not typically produce language that differentiated the city they were describing from other large or small cities in other parts of the world. Four of the samples appeared to reflect an emerging ability to avoid the pitfalls of lexical errors that obscured the intended meaning and the beginning of the ability to include more varied information.

For example, an Intermediate-High test taker produced the following description of Fez about eight and a half minutes into the test. The transcript is as follows:

طيب أنت الآن في فاس نعم؟ كيف هي مدينة فاس؟

مدينة فاز (= فاس) ممتازة انا أسكن هنا مع عائلتي المغربية لذلك هذا الصيف فرصة ممتازة لأشاهد

الحياة في في بيت مغربي حقيقي وأسكن في مدينة قديمة هذه منطقة في المدينة فاس ممتعة جدا أقدم قديم مدينة قديمة جدا مع تاريخ طويل في مدينة قديمة في فاس يوجد أقدم جامعة في العالم جامعة القروييونوهناك كثير من ورشات ودكاكين مع أشياء تقليدية ومثلا فيه هناك بدارجة دار الدبغ مكان حيث الناس يصنّعون (=يصنعون) أشياء جلدية هناك كثير من أشياء مثل هذا في فاز فاز ممتازة

So you're now in Fez, right? How is the city of Fez?

The city of Fez is great, I live here with my Moroccan family so the summer is a great opportunity to see life in a real Moroccan house and I live in an old city, this area in the city of Fez is very interesting, oldest- old- a very old city with a long history in an old city in Fez there is the oldest university in the world al-Qarawiyyin University, there are a lot of workshops and shops with traditional things, and for example in Moroccan dialect there is Dar al-Dabagh a place where people manufacture [pronunciation is unclear] leather things, there are a lot of things like that in Fez, Fez is great*

This description includes much more information than most of the other attempts at the Intermediate-High rating level. The test taker includes not only her opinion, but also some of the distinguishing places in the city like a famous university and traditional crafts area found there. She also includes the fact that she lives with a Moroccan family and is able to see life in a real Moroccan family. The test taker stumbles over "oldest" and "old" in a "very old city with a long history," but overall her attempt at description is fairly clear. The examiner does not prompt the test taker for clarification. Given that she has included some information that distinguishes Fez from other cities and largely managed to avoid lexical errors that obscured her meaning like other test takers, her attempt appears to be one of the strongest of the Intermediate-High test taker sub-samples.

Superior Rating Level Descriptions

There were two requests for city descriptions at the Superior level by E2 and both were of Alexandria, Egypt. Both have more content and more varied content than the descriptions typically produced at the Intermediate and Advanced levels. Both were also elicited earlier in their tests, with both being produced in the first ten minutes of the test takers' tests. The first of the two descriptions was elicited approximately four minutes into the interview:

حدثني عن مدينة الاسكندرية سألتني عن مدينة الاسكندرية وأنا قلتُ لكَ إني ما زرتها لذلك حدثني عنها لا أعرفها

يعني مدينة الاسكندرية يعني من يعني أجمل مدن بالنسبة لي في مصر يعني طبعاً يعني تعرف أغلبية الناس القاهرة من أي مدينة أخرى ولكن بالنسبة لي يعني الاسكندرية يعني أحسن مدينة في مصر أنا ليس عندي مشكلة في القاهرة يعني زرتها ويعني أحببتها جدا ولكن مدينة الاسكندرية يعني أكثر يعني راحة من القاهرة العاصمة يعني لأن درجة الازدحام ويعني الاحتقان يعني في الشارع يعني ليس يعني كما يكون في القاهرة في القاهرة يعني مدينة يعني مزدحمة جدا جدا والاسكندرية يعني طبعا يعني كل مدينة*؟ مصرية يعني تحتوي من يعني يعني ازدحام إلى حد ما ولكن اسكندرية يعني بجانب البحر والناس طيبين جدا يعني عندهم يعني السمك في الاسكندرية يعني لذيذ جدا جدا يعني لا أستطيع أن أقول لك يعني يجب عليك أن تزور الاسكندرية وتَجرَب(= تجرّب) في بعض المطاعم الفخمة التي يعني توجد هناك ويعني

Tell me about the city of Alexandria, you asked me about the city of Alexandria and I told you that I had not visited it so tell me about it, I don't know it

Well the city of Alexandria like is among like the most beautiful cities in my opinion in Egypt like of course like the majority of people know Cairo more than any other city but in my opinion like Alexandria is like the best city in Egypt I don't have a problem in Cairo like I visited it and like I liked it a lot but the city of Alexandria is like more like comfortable than the capital of Cairo because the degree- the degree- the degree of crowdedness and like congestion like in the street like is not not like as it is in

Cairo in Cairo in Cairo is like a very very crowded like city and Alexandria like of course like every Egyptian city like includes like like crowdedness to a certain extent but Alexandria is like next to the sea and the people are very nice like they have like fish in Alexandria is like very very delicious like I am not able to tell you like you have to visit Alexandria and try in some of the luxury restaurants that like are there and like-*

This test taker's description of Alexandria includes the following elements: 1) his opinion on the beauty of the city, 2) an implicit comparison between Alexandria and Cairo, 3) Alexandria's crowdedness, 4) geographic location, 5) a famous food from the city, and 6) an urging to the examiner to visit the city and taste this food. This content is more informative than the content produced in descriptions produced at other rating levels. His vocabulary differs from that found in other samples as he calls Cairo the capital and mentions "congestion," "luxury," and "the majority of people."

The second of the two Superior level descriptions was produced after the first 5 minutes of the test:

حدثيني عن مدينة الاسكندرية

مدينة اسكندرية يعني جميلة ولكن كثافة السكان كثير كبيرة هناك الكثير من الناس في مكان صغيّر مدينة الاسكندرية تقع على ساهل (=ساحل) البحر الأبيض المتوسط ولذلك دائما هناك جو متعدل وهناك ثقافة تشبه الثقافة في أوروبا إلى حد ما هناك التأثير من الاستعمارات من قَبل اليوروبيين (= الأوروبيين) ولذلك العمارات والمأكولات والتقاليد تأثرت بهذا النفوذ وأيضا الناس في الاسكندرية يتكلمون اللهجة المصرية وأيضا الكثير منهم يتكلمون اللغة الفرنسية أيضا بسبب تأثير كما قلت الاستعمار وماذا أيضا

وما هي أهمية هذه المدينة بالنسبة لمصر؟

المدينة مهمة للسياحة لأنّ الكثير من المصريين يجون إلى مدينة الاسكندرية في الصيف لأنّ الجو هناك كما قلت معتدلة وأيضا هناك الكثير من الآثار من العصور اليونانية من الشعب اليوناني وأيضا من الشعب الروماني وبالإضافة إلى ذلك مدينة الاسكندرية مهمة للتجارة على البحر هناك الكثير من

المراكب وهناك ميناء في مدينة الاسكندرية والميناء مهم جدا للاقتصاد وللتواصل بين الناس ولعدة أسباب أخرى

Talk to me about the city of Alexandria

The city of Alexandria like is beautiful but the crowdedness is a lot- is high there are a lot of people in a small place the city of Alexandria rests on the coast of the Mediterranean Sea and so there is always adjustable weather and there is a culture that resembles the culture in Europe to a certain extent there is an influence from the colonizations before the Europeans and therefore the buildings and foods and the traditions were affected by this influence and also the people in Alexandria speak Egyptian dialect and also a lot of them speak French also because of the effect as I said of colonization and what else?*

What is the importance of this city to Egypt?

The city is important in terms of tourism because many Egyptians come to the city of Alexandria in the summer because the weather there is as I said temperate and also there are a lot of ruins from the Greek ages from the Greek people and also from the Roman people and in addition to that the city of Alexandria is important to trade on the sea there are a lot of boats and there is a port in in the city of Alexandria and the port is very important to the economy and to the connection between people and for several other reasons

This entire exchange took approximately two minutes to complete. The test taker's description briefly includes: 1) the beauty and crowdedness of this city, 2) its geographic location, 3) its temperate weather, 4) the culture, 5) the ruins and their relationship to the history of the city, 6) the architecture, foods, and traditions affected by this history, and 7) the languages of the city's inhabitants. When the test taker asks what

else she can say, the examiner adds a related question about the importance of this city to Egypt. The test taker then states that: 1) the city is important to tourism because a lot of Egyptians go there in the summer, 2) there are a lot of ruins from the Greek and Roman periods, 3) it is important to sea trade, and 4) its port is important to the economy. Her description is also marked by her use of vocabulary and phrases that were not produced in other descriptions like "influence," "it rests on the coast of the Mediterranean," and "port." There is a minor error when the test taker mispronounces "temperate" by switching the consonants the first time and produces a word similar to "adjustable" rather than the one she intended. However, her second utterance of the word is correct, although it was pronounced as if there was incorrect gender agreement applied to it.

In summary, I found that Advanced-level speakers were able to produce descriptions that began to differentiate between cities, although often in a very limited manner. They were also able to include less personal perspective on the cities they discussed and instead include more information about the cities in question. In contrast, the Intermediate speakers were not always able to respond to the request for description in an intelligible manner. Those Intermediate speakers who did were more likely to focus on basic information about the city like whether it was big or small, whether it was crowded or not, and what the buildings and streets looked like. Their attempts typically did not produce language that could serve to distinguish one city from another, other than by general categories like size. The Superior test takers included the same information as the Advanced and Intermediate level speakers, but they were also able to add information about geography or history, both of which served to distinguish the city they were describing from other cities of similar size.

Narration

In the following section, I present the results of my analysis of the narration sub-samples, presenting a selection of the samples and beginning with the Advanced rating levels. I judged the quality of these responses by considering which speech samples could be categorized as narration according to Labov's definition. My broad finding from this analysis is that Advanced-level speakers produced narration according to Labov's definition but that lexical gaps often obscured important content in these narrations or noticeably delayed the narration. In addition, Advanced speakers often needed prompting or support from examiners in order to bring their narration to a recognizable close. In contrast, the Intermediate-level speakers were often able to produce only what Labov refers to as *minimal narration*, i.e. one temporally related juncture, and sometimes failed to include "reportable" events in their responses or even minimal narration. Finally, the two Superior rating level speakers used more varied vocabulary to meet Labov's narration conditions. This and other differences will be discussed in more detail below.

In the following section, I will focus on the Advanced level samples I found to be the strongest performances from Advanced rating level test takers. I then focus on the contrast between the Advanced and other rating level samples in order to examine in more detail the ways these groups differed in the quality of their narration. All transcriptions included in the section begin with the examiner's first request on the subject (which may or may not include a request for narration initially) to the test taker's final utterance on the same subject. Examiner contributions are underlined to differentiate them from test taker production and an English translation of the Arabic transcript is included for each sub-sample.

The highest-level sub-sample was elicited by E2 from a test taker who was subsequently rated Advanced-High. The following is from approximately 13 minutes into the test. The transcript includes question marks indicating unintelligible words, asterisks indicating words transcribed exactly as they were heard, and underlined spaces to represent personal or place names that were redacted from the transcripts:

طيب حدثني عن قصة حدثت معك وأنت في القاهرة اه ه طرفة يعني تتذكرها دائماً

قصة حقيقية يعني؟

نعم لو كان عندك قصة تحدثني إياها يعني

نعم أنا أحكيلك قصة والله والله فظيعة يوم من الأيام رحت مع أصدقاء مع ببعض أصدقائي إلى فندق وذهبنا إلى السطح عشان نشرب مع بعض ونتكلم ونستمتع بالجو الجميل فكنا نشرب وكان فيه بنت جميلة جداً فأنا قررت لا يعني أصحابي اقتعنوني إنه بأن أتكلم معها فأنا ذهبت في البداية أنا لا كنت أريد أن أذهب وأتكلم معها ولكن بعدين ذهبت وتكلمت معها وهي اعطيتني رقمها تليفون وبعد ذلك أنا ذهبت إلى البيت أنا روّحت وكنا نتكلم بتليفون شوي بعد ذلك اليوم التالي قررنا أن نلتقي مع بعض فبس أنا حسيت يعني في الليل ماكنتش ممكن أنام يعني أنا مش عارف أنام أنا كنت قلقان كتير يعني أنا حسيت أنه فيه حاجة غريبة في الموضوع دا فهمت كيف؟ فاليوم التالي أنا ذهبت إلى القهوة وشربنا قهوة مع بعض وكنا نتكلم نشرب شيشة وأنا حسيت أنه كان فيه حاجة غريبة مش ممكن أنا أشرح لك شو هذه هذا الشعور اللي أنا حسيت بس حسيت أنه كان في حاجة غريبة فبعد ذلك قالت لي _____ ليش لماذا لا نقوم بالمشي فمشينا مع بعض في منطقة حلوة اسمها _____ وكان فيه راجل تمشي* بالاتجاه العكسي فهو شافني و؟؟؟ وهو قال لي معاك شو الاسم بالفصحى ولاعة؟ تعرف ولاعة صح؟

نعم نعم نعم نعم

ولاعة مش عارف بالفصحى فهو سألني معاك ولاعة؟ فقلت له نعم عندي ولاعة وبعد ذلك أنا استمرّرت بالمشي وأنا حسيت أنه كان فيه واحد ورائيَ فهمت كيف؟ وأنا لاحظتُ وفعلاً هو كان وراني هو كان ورائيَ فأنا كنت استمررت بالمشي وتاني مرّة لاحظت ورايَ وهو موجود فالبنت اللي كانت تمشي معاي سألتني _____ لماذا أنت قلقان؟ لماذا أنت تلحظ الوراء؟ فأنا قلت لها أنا حاسس إنه هو الرجل دا هو عايز حاجة فهي قالت لي لا _____ بس امشي نمشي مع بعض

وخلاص انسي الموضوع فأنا حسيت أن الرجل دا كان عايز حاجة مختلفة وهي أنا حسيت أنه هي اتغيرت وأنا قلت لها أنه أنا أخدت بالي أنه كان فيه واحد ورائيَ فأنا قلت لها لماذا لا نذهب مع بعض ونتعشى في مكان تاني فركبت تاكسي معاها ولما كنا في التاكسي السواق هو شافني زي يا راجل فيه حاجة غريبة وهو عارف المدينة يعني هو سواق من زمان فلما شفت الكلام دا وهي اخذت الموبيل وكانت تمسج حد أنا مش عارف مَن هي تمسج فهمت كيف؟ فأنا كلمت صاحبي وقلت له خلينا نلتقي مع بعض عشان أنا وأنا بأتكلم (لغة أخرى) معه عشان فيه حاجة غريبة أنا بامشي مع بنت ولاحظت أنه فيه كلام غريب يعني فهو قال لي ماشي أوكي مافيش مشكلة خلينا نلتقي في منطقة (ثانية) فلما وصلنا إلى (المنطقة) أنا شفت صديقي ورحنا لـ (مطعم) وهي قالت أنا عايزة اروح التوليت فقلت لها اتفضل* اتفضّلي مافيش مشكلة أول ما هي دخلت التوليت أنا وصاحبي جرينا!

Well, tell me a story that happened while you were in Cairo uh, an interesting one that you'll remember always

A true story you mean?

Yes, if you have a story, tell me what happened to you

Yes, I'll tell you a story that's you know really crazy, one day I went- I went with some of my friends to a hotel and we went to the roof so we could drink together and talk and enjoy the beautiful weather so we were drinking and there was a very beautiful girl there so I decided no like my friends convinced me that to talk to her so I went and in the beginning I didn't want to go and talk to her but after that I went and talked to her and she gave me her phone number and after that I went home I returned home and we were talking on the phone a little after that next day after that next day we decided to meet together so- but I felt at night I couldn't- it wasn't possible to sleep you know, I couldn't sleep, I felt very worried and I felt that there was something weird about this situation, you know what I mean? So the next day I went to the coffee shop and we drank coffee together and we talked and smoked shisha and I felt that there was something weird I can't explain to you what these feelings were that I felt but ah I felt that there was something weird so after that she said to me _____ why don't we go for a walk? So we walked together in a nice area called _____ and there was a man*

walking in the opposite direction so he saw me and said to me "Do you have" ah what is the name of it in formal Arabic? Lighter? You know lighter, right?*

Yes, yes, yes, yes

Lighter- I don't know what it is in formal Arabic he said to me: "You got a lighter?" and I said to him yes I have a lighter, after that I continued to walk and I felt that there was someone behind me you know what I mean? So I felt that there really was someone behind me and I was walking and again I ??? behind me and he was there so the girl who was walking with me asked me why are you worried? Why are you ??? behind? So I said to her I feel that that man he wants something so she said to me no _____ just walk! We're walking together and that's it, forget about it! So I feel like this man he wants something different and she I felt that she is changed so I told her it has caught my attention/ is bothering me that there was someone behind me so I said to her why don't we go together and eat dinner somewhere else? So I took a taxi with her and when we were in a taxi the driver he watched at me like, "hey man, there's something weird here" and he knows the city and he has been a taxi driver for a long time so I saw this talking* and she had taken her cell phone and was messaging someone I don't know who she was messaging you know what I mean? So I talked to my friend, "Let's meet together because-" and I'm talking in [another language] with him "because there's something weird, I'm walking with the girl and I notice there is some weird talk*" so he said to me okay he said to me no problem, let's meet in an area called _____ so when we arrived at _____ I saw my friend and we went to [a restaurant] and she said I want to go to the bathroom okay, go ahead*- go ahead no problem, as soon as she entered the bathroom, my friend and I ran and left that place*

This test taker's production lasted approximately four minutes with little input or interruption from the examiner. This Advanced-High test taker satisfies Labov's narration requirements by providing an orientation (we were drinking … there was a pretty girl), a complicating action (I felt there was something weird), and a resolution (we ran and left that place). He accomplishes this with no prompting or assistance from the examiner, a fact that also distinguishes his performance from test takers at lower levels.

Although the production is mostly clear, there are parts of this test taker's speech that are either obscured by mispronunciation or grammatically inconsistent with what he appears to be communicating (for example, when it sounds as if he is using the wrong conjugation for "the man is walking"). In addition, there is a lot of repetition in this narration. The speaker may be using some of it to build suspense or add emphasis or possibly to include the listener in his story; however, some of the repetition appears unnecessary, as the point has already been clearly made to the listener (in particular the repetition of "a strange thing").

Advanced-Mid Narrations

From this narration at the Advanced-High level, I will now move to the sub-samples elicited at the Advanced-Mid level. In what follows, I will discuss three of the strongest examples from this sub-level.

The one of the most detailed examples of narration was elicited by E2 and occurred after approximately 18 minutes. Prior to this exchange, the test taker and examiner had been discussing the political unrest in Syria and the test taker is referring to this when she mentions "problems."

طيب وأنت في دمشق أو وانت في الأردن يعني حدثيني عن أغرب قصة حدثت معك

لازم أفكر في ذلك السؤال أخذت رحلة حول سوريا عندما كنت ساكنة بسوريا واستأجرت سيارة في دمشق وقدت* سيارة حول سوريا في شهر آذار آذار في بداية المشاكل وفي ذلك الوقت المشاكل بدأت في درعا ولكنني نويت أن أن آخذ هذه الرحلة فذهبت وكان هناك مشاكل مع لواء الأمن في في عدة مدن في سوريا لأنني كنت أجنبية في سوريا في وقت خاص ومثلاً لما كنت في مدينة صغيرة في جبال سوريا كان كنت ماشية في الشارع وكان هناك رجل يتكلم معي عن أو يسألني عن ما* أنا واتهمني بجسوسة وكنت خايفة جداً وبُعَيدَ ذلك الناس من الشارع يُلطقِني* وضربوا على سيارتي ولحقوني* (=لاحقوني) خارج المدينة

أذاً كيف خلصتِ من هذا الموقف يعني هربتِ؟

نعم هربت الحمد لله

الحمد لله أذاً كانت غريبة وخطيرة

وخطيرة نعم

ماذا تقول لك عن العقلية السورية؟

انهم خايفين من الخارج من المستقبل من أي شيء غريب يخوّفهن

So, while you were in Damascus or while you were in Jordan, you know, tell me the strangest story that happened to you.

I have to think about that question I took a trip around Syria when I was living in Syria I rented a car in Damascus and I drove a car around Syria for the month of March March in the beginning of the problems and at that time the problems started in Daraa but I wanted to to take this trip so I went and there were problems with military heads of security in various cities in Syria because I was a foreigner in Syria in a special time and for example when I was in a small town in the mountains of Syria it was- I was walking in

the street and there was a man talking to me about- or he's asking me about what I am and he accused me of spying being a spy and I was very scared and after* after* [remoteness?] that* people from the street gathered* and hit my car and followed me outside the city*

Then how did you get out of this situation, did you flee?

Yes, I fled, thank God

Thank God, then it was strange and dangerous

and dangerous yes

What does it tell you about the Syrian mentality?

That they're afraid of the outside, of the future, of anything strange that scares them

This exchange lasts approximately four minutes, as the test taker is an extremely careful speaker whose delivery is somewhat slower than other test takers in this data set. However, her production here is one of the most detailed found at the Advanced-Mid level and she does not evince any difficulty with words or expressions. In terms of content, she includes several key elements, explaining: 1) that she rented a car, 2) that she spent the month driving around the country, 3) that she was traveling while political unrest was breaking out, and 4) that she had problems getting permission to travel in different cities. All of this content provides a rich orientation to the complicating event that she relates. After presumably securing permission to travel in these areas, she says that she ran into a man who accused her of being a spy, and that some people gathered and banged on her car. At this point, the examiner prompts her to add a resolution to the story and she reports that she fled, using the same word the examiner used in his question. He also prompts her for an evaluation, stating that the story was not only weird, but also dangerous, an evaluation with which she agrees. Lastly, he asks for a more

abstract evaluation of the Syrian mentality and she responds that she thought these people were afraid.

This is one of the stronger examples of narration at the Advanced level because there are no examiner follow ups that indicate that the speech was unclear; instead the examiner questions are focused on expanding or assessing the events the test taker has narrated. However, it is noteworthy that the examiner essentially provides the coda to the story by suggesting that the event was not only weird, but also dangerous. Likewise, the examiner suggests a conclusion to the narrative and uses the word "fled" in his question. Supplying this conclusion certainly indicates the examiner's interest in what the test taker is saying and may have served to put the test taker more at ease, which in turn could have produced a better speech sample overall. However, in terms of the evaluation of this sub-sample, it is unclear as to whether or not the test taker knew and understood these words or simply accepted the evaluation given by the examiner in order to bring this topic to a close.

In contrast, the second narration sample represents a complete narration that is only very slightly obscured by lexical gaps. It was elicited by E1 and occurs approximately 7 minutes into the interview. The transcript is as follows:

أنت يا _____ سافرت إلى بلاد مختلفة هل يمكنك ان تحكي لي قصة غريبة أو ظريفة حدثت لك في هذه الاسفار الكثيرة في البلاد المختلفة؟

نعم يا أستاذ عندما وصلت إلى الأردن لأولى مرّة لم أعرف أي شيء عن العربية العامية فكنتُ عرفتُ اللغة العربية الكلاسيكية الفصحى فقط وعندما وصلتُ وخرجتُ من المطار كنتُ في شارع أمام المطار مع كل شنطاتي وإلى آخره فشاهدتُ واحد من الشوفير السائقون السائقون التاكسي خارج المطار فسألته سألت له هل من الممكن يا حضرتك أن تذهب معي إلى منطقة اسمها جبيها يعني هذا كلام غريبة وشاهدني الشوفير التاكسي كأني من من القمر حتى و"هل أنت مغربي؟ أو... "مين أنت وأيش بدك؟" وهذه الـ... هذه القصة تمثل الدغلوسية؟ أنا لا أعرف كيف أعبر عن هذا بالعربي هناك لغتين أو ممكن ثلاث لغات : اللغة الكلاسيكية والعامية الاردنية ولغة النسان؟ الانسان؟ كيف نقول

teeth?
[clicks]
" لا أريد هذا " هذا صوت

You_____ traveled to different countries, well, can you tell me a strange or nice story that happened to you in these travels in various countries?

T: Yes, when… when I arrived in Jordan the fairst time I did not know anything about Arabic dialect since I had known Classical Fusha Arabic only and when I arrived and I left the airport I was in a street in front of the airport with all my suitcases and all that and I saw one of the drivers drivers taxi drivers outside the airport so I asked him asked to him "Would it be possible O sir for you to proceed with me to a neighborhood called Jubaiha?" you know, that was strange speech and the taxi driver looked at me as if I was from … from the moon almost and: " Are you Moroccan or …?" "Who are you and what do you want?" and this … this story shows the diglossia? I don't know how to express this in Arabic. There are two languages or maybe three languages … the classical language and Jordanian dialect and the language of ??? the person? How do we say* ***teeth****? "I don't want that" That sound*

This test taker only requires one prompt to produce a fairly robust narration. The test taker provides orientation ("when I arrived in Jordan for the first time…I left the airport"), complicating actions ("I asked a cab driver … he looked at me like I was from the moon"), and evaluation ("that was strange talk!"). He also uses a dialect equivalent for the word "driver" and then produces the more formal word, indicating that he knows these terms differ in register.

However, the test taker encounters some difficulty when he tries to expand his evaluation of the story. This starts when he tries to jokingly suggest that Jordanians speak their dialect, use formal written language, and also communicate with gestures and noises made with their teeth. He is unable to find a translation for diglossia so he Arabizes its English pronunciation. He also attempts to say that some people communicate by making noises. He calls this "a language of teeth," producing two words that appear to be approximations of the word "teeth" until he ultimately appeals for the examiner's help by using English. In a non-testing context, this strategy would most likely be a sound one, given that the test taker cannot recall the word correctly. However, the examiner is prevented from helping him because he is administering a test to the test taker. Therefore, the test taker's final comments about the story can only be fully understood by a listener who understands the English for "diglossia" or "teeth," or by a listener who surmises that the test taker is referring to a nonverbal method of saying "no" by raising one's eyebrows and making a clicking sound with one's teeth.

The third narration sample provided by an Advanced-Mid speaker was elicited by E9 after approximately 8 minutes:

خلال عملك في هذه المنظمة الخيرية هل تذكرين اي قصة حدثت معك ممكن أن تخبريني إياها؟

بسبب حقوق السَريّية يعني بالنسبة للعلاقة بين المحامية وبين الزباين ولكن أتطوع بخاص أتطوع مع الزباين الذين لديهم مشاكل بالنسبة للاعانة الأطفال ودعم الأطفال ... ا ه ه أحاول أن أفكر عن قصة مناسبة ولكن للأسف كل القصص فيها الحزن وكلها يعني ليس لديها نتيجة ا ه ه ! ولكن هذا قبل شهرين كان لديّ زبونة وكانت تخاف من أب طفلها لأن ظنت أن هو أراد أن يترك البلد مع الطفل ليذهب إلى بلده من خارج أمريكا وكانت تخاف من ذلك فلذلك أنا كتبت رسالة إلى وزارة

خارجية الامريكية لأطلب لها أن لو الأب قد حاول أن يحصل على جزء* السفر جواز السفر للطفل ليست من الامكانية أن وزارة الخراجية تمنع تمنعه من ذلك لأن هو كان أب الطفل وكان حقه ليفعل ذلك ولكن كان من الامكانية أن الوزارة الخارجية أن تعلنها* (=تعلمها) لو الأب حاول أن يحصل على جواز السفر للطفل لكي الأم تعرف أن ممكن هذا قد يعني الأب سوف يحاول أن يترك البلد مع الطفل وهذا كان شيء سهل بالنسبة لي ولكن أظنّ هذا كان شيء مفيد وجيد جداً بالنسبة للأم لأن هذا أعطيت عليها يعني سلامة العقل وإمكانية أن تسيطر على الحالة نوعاً ما

So during your volunteer work with this charitable organization do you remember a story you can tell me?

because of the privaycy rights you know in terms *the relationship between the lawyer and the clients but I volunteer particularly I volunteer with the clients who have problems in terms of child support … ah I'm trying to think of an appropriate story but unfortunately all the stories are sad and all of them don't have a result … ah oh! But this two months ago I had a client and she was afraid of her child's father because she thought that* he wanted to leave the country with the child go to his country from outside America and she was afraid of that so as a result I wrote a letter to American *Department State so I [could] ask for her if the father had tried to obtain a piece- a passport for the child …the…the…it is not possible for the State Departments to forbid him that because he was the child's father and it was his right to do that but it was possible for the State Department to advertise* [notify] her if the father had tried to get the passport for the child in order that the mother know… meaning that the father will try to leave the country with the child*

This test taker's narration from her volunteer work lasts approximately two minutes. While her delivery and speech are not perfect, her story contains all the elements required by Labov's definition. First, she provides an orientation as she explains the clients she worked with. Second, she provides a complicating action concerning a particular woman and the woman's concerns about her child's father possibly obtaining a passport for their child without her knowledge. Third, the test taker provides a resolution that signals that she is done with her narration by noting that she was able to secure a letter from the Department of State which addressed the mother's concern.

In addition to producing language that meets the requirements of narration, this test taker further displays an ability to respond to the examiner's prompt without actually beginning a story. Her initial utterances allow her to demonstrate her willingness to respond while she searches her memory for an appropriate story; while she is doing this, she comments on how these stories are very sad and have no "result" (or resolution). However, she does say "advertise her" when she means "inform her," a fact that the examiner did not comment on. There is also one instance in which the test taker tries to say "passport," but initially says "piece."[29] The test taker corrects this mistake without any examiner reaction.

The third narration example from the Advanced-Mid level was elicited by E9 at 3:40 in the interview and is the earliest request for narration found in this data set. This coupled with the fact that the examiner cut off the test taker's speech multiple times during the interview leads me to treat this sample with some reservation. The quality of the recording is the same as other samples so it is unclear what was causing the examiner to cut off this particular test taker's utterances with regularity; E9 did not exhibit this

[29] These words share two letters in Arabic, i.e. "j" and "z," and both begin with the same letter. However, they are otherwise unrelated.

behavior with other test takers. However, this is the weakest narration among the Advanced-Mid samples. The transcript of the exchange is as follows:

طيب خلال سفرك لنيجيريا هل تذكرين أي شيء قصة شيء حدث معك تخبريني إياه؟

...آسفة أنا ما سمعت ال

عندما كنتِ في نيجيريا

نعم

هل حدث معكِ مثلاً قصة شيء مضحك حادثة؟

اااه مضحك

مضحك نعم غريب

اااه غريب

اخبريني عن قصة حدثت معك عندما كنت في نيجيريا

ممكن شيء غريب ا ه ه بعض الأكل كان غريب شوية

أخبريني ماذا حدث

انا ذهبتُ إلى المطعم

Nigeria

ومع ال الموظفين في المنظمة وطلبتُ منهم ل ل طلبتُ الأكل* يعني روز*(=رز) وفول وهكذا ورأيتُ أكل أخرى* وحاسة* (=خاصة) وهذا كان الجلد جلد من ال لا أتذكر ال ال لا أتذكر ب ب ب الظبط أتذكر

جلد حيوان؟

الحيوان نعم الجلد الحيوان فقط مع بيض ونعم بعض التوابل وكان غريب شوية غريب شوية

نعم و هل تذوقت هذا الطعم؟ كان جيد؟

لا لا انا اخفتُ

نعم

نعم

So during your trip to Nigeria do you remember anything a story of something that happened to you that you can tell me about?

Sorry I didn't hear the …

When you were in Nigeria

Yes

Did anything happen to you for example a story or something funny

Ahh funny

Funny yes weird

Ahh weird

Tell me about a story that happened to you while you were in Nigeria

Maybe something weird- some of the food was a little weird

Tell me what happened

I went to the restaurant ***Nigeria*** *and with the employees in the organization and I ordered from them I ordered food like rice* and beans and that [sort of thing] and I saw other special* food and and this was the skin skin from the- I don't remember the I don't remember the the I don't remember exactly I remember*

Animal skin?

Animal yes the skin the animal only with eggs and yes some spices and it was a little weird a little weird

Yes and did you taste the food? Was it good?

no no I was afraid

Yes

Yes

The test taker's responses form what Labov refers to as the "skeleton of a narrative" (Labov, 1972, p 361). Her first response ("the food was a little weird") serves as an abstract for what follows. The test taker then produces clauses that are temporally related and reports encountering the skin of an animal being served in a restaurant. This is a reportable event as the examiner asked for a weird or funny story, and the test taker's abstract indicates that she regarded this food as unusual. Thus, the test taker has fulfilled the minimum requirements for narration, although her production is not as clear as other test takers' attempts and it lacks an evaluation. The examiner finally prompts her to produce one by asking, "Did you taste it?" and the test taker closes by stating that she was afraid to try this food.

Advanced-Low Narrations

In contrast to the Advanced-Mid speakers, the Advanced-low speakers appeared to struggle more in responding to narration requests. There were three requests for narration at this level, two of which elicited narration attempts. There was also one test taker who volunteered narration. I will present these three samples in turn to examine the ways in which these test takers were more or less successful.

A test taker volunteered one of the clearest narration samples after about 17 minutes with E2. This exchange took approximately two minutes:

طيب وماذا تفعل عادة في نهاية الأسبوع؟

في نهاية الأسبوع طبعاً طبعاً أدرس دائماً في هذا البرنامج دائما واجبات كثيرة أنا أفعل اشياء كثيرة مع أصدقائي مثلاً من أسبوعين أنا أنا سافرت إلى حفلة زفاف ا ه ه أقارب شريكي في السكن شريكي في السكن اسمه ــــــــــــ وأنا سافرت إلى طيب القصة هو هذه القصة هو قال لي من اللازم أن نروح أو من اللازم أن نسافر إلى حفلة زفاف في عائلتي وأنا قلت واو هذا جميل طبعا أنا أتفق أنا أريد أن أسافر في في يوم جمعة سافرنا إلى بيت عائلته في ضواحي ضواحي اربد والجلسة كان* جميل كان مؤكولات لذيذة وحوار جميل وكل الشيء مثل ذلك ولكن ما كان فيه أي دليل من أي دليل الزفاف ما شفت العرس أو العروس أو العريس أو أي شيء مثل هذا لكن

نعم

بغض النظر عن ذلك كان كان يوم ممتع كثير أو ممتع جداً

Ok and what do you normally do on the weekends?

On the weekend of course of course I always study in this program, there is always a lot of homework, I do a lot of things with my friends for example two weeks ago I I traveled to a wedding party uh for relatives of my roommate my roommate's name is _____ and I traveled to- well, the story is this story- he said to me, it is necessary that we go it is necessary that we travel to a wedding party in my family and I said "Wow, that's great" of course I agree I want to travel, on on Friday we traveled to his family's house in the suburbs suburbs of [the city] and the event was beautiful the food was delicious and nice conversation and everything like that but there wasn't any evidence of a wedding I didn't see the wedding or the groom or the bride or anything like that but*

Yes

Besides that, it was it was a very interesting day, or very interesting[30]

In this narration, the test taker supplies an orientation ("last weekend, I traveled"), a complicating action ("no evidence that it was a wedding"), and an evaluation ("besides that, it was very interesting"). This is one of the strongest examples at this level as the test taker supplies a complete narration without prompting or clarification from the examiner and likewise exhibits no lexical gaps that could cause misunderstanding. Also worth noting is the fact that his use of the connector "besides that" allows him to smoothly add his evaluation to his other statements about his trip.

The following narration attempt was elicited by E9 approximately 7 minutes into the test. The word in bold was spoken in English:

ممتاز طيب وخلال تواجدك في المغرب أنتِ تزورين مدن كثيرة ومناطق كثيرة هل تذكرين أي قصة حادثة أو شيء اخبريني إياه مثلاً شيء حدث معكِ كان غريب أو مضحك

اااه نعم عندما زرتُ مدينة طنجة زرت زرتُ البرنامج _____ في طنجة يعني عندي صديق في هذا البرنامج وذهبنا إلى الشاطئ وانا فتُ* في البحر مع نظراتي نظاراتي فقط ؟؟؟ لل بحر فـ فقدتهم فقدتها

نعم

وهذا كانت مشكلة كبيرة لأنّ لم أرَ أي شيء كل اليوم ومن اللازم أنْ كان من اللازم أنْ أسافر إلى فاس هذا اليوم و

نعم

نعم انا كنتُ في في ال
Train
وكان ليس عندي نظارات

[30] This test taker says "very interesting" (*mumtʿa kathiir)* one way before supplying a restatement using another word for "very" (*mumtʿa jiddan)*.

وضع صعب نعم نعم

نعم نعم

وضع صعب وماذا فعلتِ بعد ذلك هل اشتريتِ نظارات جديدة؟

االاه نعم اشتريتُ نظارات جديدة عندما ارجع رجعتُ إلى فاس ولكن كانت صعب هذا كانت صعبة *أيضاً لأنّ الورقة

مع ؟؟؟ الرقم لل* نظاراتي كان أمريكية فالطبيب لم لا* يقرأ الرقم صحيح* ولذلك من اللازم أنْ أرجعُ يوم أخرى لل ف االاه نعم

هذا وضع صعب جداً ولكن يحدث

نعم

عندما نكون في الخارج نعم

وغالي جداً

غالي جداً نعم بالطبع بالطبع

Great, ok, and and during your time in Morocco you're visiting a lot of cities and a lot of areas, do you remember any story or event or anything for example that happened to you that was strange or funny?

Ah yes when I visited visited the _____ program in Tangiers you know I have a friend there in the program and we went to the beach and I went into the sea with my glasses my glasses only and ??? in the sea I lost them lost them*

Yes

And that was a big problem because I didn't see anything all day and it was necessary and it was necessary that I travel to Fez that day and

Yes

Yes and I was in in the ***train*** *and I didn't have my glasses*

A difficult situation yes yes

Yes yes

A difficult situation and what did you do after that, did you buy new glasses?

Yes, I bought new glasses when I am returning I returned to Fez but it was hard it was this was hard too because the paper with ??? the number for my glasses was American so the doctor didn't doesn't read it correctly and as a result it was necessary that I go back another day to- yes*

That is a very difficult situation but it happens

Yes

When we're abroad yes

And very expensive

Very expensive yes of course of course

This exchange lasts approximately 2 minutes and 40 seconds. The test taker successfully provides an orientation ("I visited … Tangiers") and a complicating action ("we went to the beach and I lost my glasses"). She further reports that she needed to travel that day and could not see anything. However, she is unable to retrieve the word for train and instead produces it in English. There are also other words that her pronunciation render difficult to decipher like "went," "went into the sea," and "paper." In addition, the examiner has to prompt her to continue and conclude her narrative, which the test taker is able to do in a simple manner. From her responses, it is clear that she was

able to get new glasses, but it is unclear what she did to fix the issue of her eyeglasses prescription being misinterpreted nor does she supply any details of how she got from the beach to the train and then home again. The parsimoniousness of this narration makes it seem less robust than the one in the preceding sample.

The last narration sample at the Advanced-Low level was elicited by E9 approximately 11 minutes into the test. The exchange was as follows:

نعم أحكي لي عن مغامرة قمتَ بها خلال السفر

عن عن ما؟ مرة أخرى؟

مغامرة

مغامرة؟

نعم

يعني القصص؟

القصة خلال السفر يعني شيء حدث معك

انا أحلى

خلال السفر

أحلى رحلتي كان إلى اليونان في وزرتُ إلى* الجزائر اليونانية و

إلى ماذا أين؟

إلى إلى اليونان

نعم وماذا زرتَ في اليونان؟

زرتُ إلى* الجزائر اليونانية وكنتُ بجانب البحر وكنتُ في فندق مع منظر على البحر جميل جداً

نعم إحكي لي عن شيء حدث معك تتذكره دائماً

شيئاً*؟

نعم شيء معين يعني موقف معين تتذكره دائماً من تجربتك

االاه يعني

مثلاً في المطار حدث شيء أو في السوق

االاه نعم نعم أفهم أفهم

شيء حدث معك

في الحقيقة الأمر حياتي مملة لا دقيقة واحدة أريدُ أنْ أفكر قليلاً

خلال السفر أشياء يعني تحدث دائماً غير متوقعة خلال السفر

في السفر أحبُ

لا ليس ماذا تحب يعني مثلاً انا تأخرت طائرتي الشهر الماضي وبقيتُ طوال النهار في المطار

في المطار والله؟

نعم هل حدث معك شيء غير ؟؟؟ ما كنتَ لم تكون تتوقعه؟

أيه كنتُ مرةً واحدةً في كنتُ في طائرةً قليلاً آسف آسف كنتُ في طائرة صغير

نعم وماذا حدث؟

وكنتُ أطور* كنتُ أطور في في الولايات المتحدة ولكنّ كان ال كان الطقس غير جيد وكان ال انتهى* ال الطيور* إلى مطار أخرى وثم انتظرنا حوالي ساعة ثم رجعنا إلى السماء وذهب يعني حدث كان هناك غائم* والريح الريح كان شديد جداً

نعم

يعني في السماء

نعم وكيف تصرف الركاب في الطائرة؟

ما الركاب؟ لا أعرف هذا هذا الكلمة هذه الكلمة

نعم يعني أنت والناس في الطائرة

كنتُ مريض كنتُ مريض وكنتُ كنتُ أجلس أجلسُ في مقعدي وكنتُ قلتُ لنفسي أستطيعُ أنْ أنْ أجلس فيها ولا أريدُ أن أصبحُ مريض جداً والحمد الله ما حدث أي شيء

نعم طيب طيب

Tell me about an adventure you had during your travels.

About what? One more time?

An adventure

An adventure?

Yes

Meaning stories?

A story during [your] travel, meaning something that happened to you

I- the most beautiful

During [your] travels

My most beautiful trip was to Greece in- and I visited the Greek islands and-

To what? Where?

I visited to the Greek islands and I was next to the sea and I was in a hotel with a view on the sea- very beautiful

Yes, tell me about something that happened to you that you'll always remember

Something?

Yes, something particular, you know, a particular situation you'll always remember from your experience

Ah like-

For example in the airport something happened or in the market

Ah yes yes I understand I understand

Something that happened to you

Actually, the thing is my life is boring no, just a minute, I want to think a little

During travel, things you know unexpected things always happen when we travel

When traveling, I like to-

No, it's not what you like you know for example I- my plane was late last month and I spent the whole day in the airport

In the airport, really?

Yes, has something happened to you that was un- that you weren't expecting?

Yes, I was one time I was one time in a plane a little sorry sorry I was in a small plane*

Yes and what happened?

And I was flying in the United States but the weather was was not good and the end of the birds [flight] was to another airport and then we waited about an hour and then we returned to the sky and went*- you know it happened there was cloudy* and the wind the wind was very severe*

Yes

Like in the sky

Yes, and how did the passengers behave in the plane?

What is passengers? I don't know this this this word

Yes, it means you and the people in the plane

I was sick I was sick and I was- I was sitting I was sitting in my seat and I had said to myself I can I can sit here and I don't want to become very sick and thank God nothing happened

Yes, good, good

This exchange lasted three minutes and 40 seconds, making it longer than the other Advanced-Low samples. However, the production from the test taker is limited as he struggled to follow the examiner's choice of words. In addition, he uses a plural of the word "island" that may have temporarily confused the examiner. The recording of the word and his pronunciation appear to be clear, but the plural is a less common one that coincidentally is also the same as the modern name for Algeria, which may have caused the confusion. Once the examiner has confirmed what the test taker is saying, the test taker describes a trip he took to the Greek islands. The examiner then interjects and continues her efforts to elicit a narration. After some negotiation, the test taker understands what is being asked of him and asks for some time to think. As a result, the actual narration attempt lasts only one minute.

The test taker is able to supply an orientation ("I was in a small plane") and a complicating event ("the weather was was not good …. we waited an hour … I was sick"). He is also able to provide a conclusion when he ends by saying that nothing happened. However, some of this information is obscured because of his substitution of birds (*Tuyyuur*) instead of flight (*Tayaraan*) early in his narration. The sample is also less clear than others because the examiner made several word choices like "adventure" and "passengers" that appeared to confuse the test taker momentarily. It took the examiner a while to request a narration in a way the test taker could understand so this sample is harder to interpret. It is possible that the test taker was not able to provide a complete

narration, but it is also unclear whether or not he could have provided a better one with a clearer solicitation from the examiner.

Intermediate Narrations

As stated above, there were 12 attempts at narration at the Intermediate level, with Intermediate-High test takers producing eight and Intermediate-Mid test takers producing four respectively. Of these, seven did not include reportable events or chronologically dependent events and therefore did not meet the requirements of narration or minimal narration; I considered these to be failed attempts. Of the remaining five sub-samples, two attempts produced skeleton narrations, which included chronologically dependent clauses but no reportable events. The remaining three sub-samples met Labov's basic definition of narration by including reportable events.

In this section, I present two of the narration attempts that met the basic requirements of Labov's definition first, one from the Intermediate-High and one from the Intermediate-Mid rating level. Second, I examine a sample each from the skeleton and failed narration attempts, in order to show the ways in which some test takers' responses did not meet Labov's requirements of narration.

An Intermediate-High test taker produced the first narration. Three question marks indicates an unintelligible word and empty spaces indicate redacted names:

طيب وأنت في عمان هل حدث معك قصة غريبة أو مضحكة؟ هل حدث معك شيء غريب؟

قصة غريبة ؟

مثلاً عندما كنتَ في سيارة الأجرة أو في الشارع؟

ممكن يعني في الشارع انا وصديقي اسمه ___ نحن مشينا في الشارع وتكلمنا مع الأردنيون عن أي شيء وهذا الأسبوع الماضي نحن تكلم* مع واحد من الأردنيون عن شهر رمضان وصديقي ___

هو قال إلى إلى الأردني يعني انت سام* والأردني هو لا أو ما فهم و___ هو لا لا عرف عرف الكلمة السام (=سَم) مختلف من صوم صوم في رمضان سام (=سم) هذا مش مشروبة سيئ سيئة

نعم

وهذا هذا مضحك في بعدَ* شوية ولكن في في نفس الوقت الأردني هو ليس ليس مضحك

نعم لأنّه لم يفهم ماذا اراد ___ أنْ يقول

Ok so when you were in Amman, did any weird or funny story happen to you? Did anything weird happen to you?

Ah a weird story?

Like when you were in a taxi or on the street

Ah I was in the street ah my friend named _____ and I, we were walking in the street and we talked with Jordanians about anything and last week I was- we talked to one Jordanian ah about the month of Ramadan and ah my friend _____ he said to the Jordanian you know, you're poison and the Jordanian he's not- or he didn't understand and [my friend] he didn't know I didn't know the word "saam" ("sam" or poison) is different from fasting fasting in Ramadan "Saam"(=poison) is not- is a bad drink

Yes

And that was funny a little while later but at at the same time [to] the Jordanian it is not not funny*

Yes because he didn't understand what [your friend] wanted to say

In this exchange, the test taker's initial response lasts approximately one minute and 40 seconds as he spends more time thinking and adjusting what he is saying than

Advanced-level speakers. His attempt meets the minimal requirements of narration as he provided an orientation ("we were walking in the street … we talked about anything with Jordanians") and a reportable event ("my friend said, you're poison?"). He also provides a clear conclusion as he explains that his friend was trying to refer to fasting and instead seemed to indicate a "bad drink." He also attempts to provide an evaluation, although the evaluation is less clear than his preceding speech. From the fragments of his utterances, we can assume that he meant that it is funny now, but at the time it did not seem funny.

The Intermediate-Mid test taker's narration attempt occurred in the context of discussing the language program the test taker and examiner had both participated in during the previous summer:

جميل جداً طيب طبعاً يعني نحن كنا نسكن يعني كنتَ تسكن مع الأساتذة ومع الزملاء ويعني نأكل وتأكلون معاً وهذا طيب هل حدث موفق غريب أو شيء كوميدي أو شيء ظريف أو شيء يعني عجيب وما زلتَ تتذكره جيداً؟

االاه

مثل قصة شيء حدث لكَ في (البرنامج)

شيء هذا كانت غريب ما ممكن مرة ثانية السؤال؟

نعم يعني يعني هل حدثت عندما كنتَ في (البرنامج) في (البرنامج) في (البرنامج)هل حدثت مشكلة مع الأساتذة أو حدث شيء كوميدي أو شيء غريب أو كان هناك طالب أو طالبة أو أي شيء مثل هذا تعرف يعني أحياناً تحدث هذه الأشياء أريد أنْ تحكي لي قصة قصة شيء حدث

االه اوكي ممكن أظن* معظم الوقت كانت* جيد جداً ولكن أظن أن ليسوا ليس هناك مشاكل مع أستاذة أو أستاذي وممكن أظن ال ال قليلاً من ال قليلاً من الزملاء االاه ليس معظم ولكن هم كانوا ليس كيفَ نقول مثل صديق و ولكن عندما كنتُ في في بداية الدراسة العربية انا ليستُ* أنْ االاه أن أتكلم أو أو أستطيعُ أستطيعُ أن أتكلم جيداً بالعربية واحد مرة عندما كنتُ في في المطعم في المدرسة االاه انا أجلس أجلستُ أو جلستُ جلستُ في ال في المطعم واردتُ أنْ أتكلم معهم وهم االاه يألكلوا تأكلوا تأكلوا االاه انا الذي االاه هم يأكلون انا ليستُ أنْ أنْ أعرفهم أنْ أعرفهم أي شيء بالعربية وهم يقولون االاه انا في في ال انا في الصف المتقدم وأنتَ في الصف الأول وأظن أنتَ ل أنتَ ليس جيد بالعربية لا أريد

أتلكم معكَ لأنّ انا ليستُ انا متقدم وهذا هذا كانت التجربة غريبة بالنسبة لي ولكن معظم الناس هم الناس من الصف الثاني صف الثالثة وصف الأربع* هم يقولون أهلاً وسهلا إذا تريد أنْ أتكلم* بالعربية معني* ااه يلّا وانا علمتُ كثيراً من منهم وكانت* الشخص أو شخصين الذي يقولون انا لا أريد لا أريد هذا كان هو

هذا غريب فعلاً يعني

يعني هذا كان غريب و

ولكن معظم الطلاب كانوا لطفاء

نعم

تمام معظم الطلاب كانوا يعني يحبون المساعدة وهذا تمام ااه طيب

So of course, you know, we were living-, you know, you were living with the teachers and with the colleagues you know we eat- you eat together and that's good, did anything happen, a weird situation or something funny or something fun or you know weird that you still remember well?

Uhh

Like a story that happened to you in [the program]?

Something weird that what- the question one more time?

Yes, you know, did something happen to you when you were in [the program]? Did a problem occur with the teachers or something "comedic" or something weird or was there a [male] or [female] student or anything like that you know sometimes those things happen I want you to tell me a story something that happened

Ah, okay, maybe, I think most of the time was really good but I think that there aren't isn't any problems with the teachers or my teachers and maybe I think a little with my classmates not most- they were like how do we say? like a friend and but when I was*

at the beginning of the Arabic studies I was not to talk or or able to speak able to speak well in Arabic and one time when I was in the cafeteria in the school ah I'm sitting- I sat in the cafeteria and I wanted to talk to them and they are eating- you all are eating- you are eating- and ahh I- that- and I don't know them I don't know them anything in Arabic and they say ahh I am in the advanced class and you are in the beginner's class and I think you* to- you are not* good in Arabic I don't want to talk to you because I am not I* advanced and that was a strange experience for me but most of the people most of the people from the intermediate class the advanced class and higher they say you're welcome if you want me to speak Arabic with me* ah let's go! and I learned a lot from from them and it was* the person or two people that* are saying I don't want- I don't want- that was it*

That's really weird you know

You know that was weird and-

But most of the students were nice

Yes

Good, well, most of the students liked helping and that's good ahh, well

This test taker strives to include a reportable event, i.e. that a few students did not want to talk to him because they felt his ability was beneath theirs; he also agrees with the examiner's evaluation that this was strange. In addition, this test taker provides more language than other test takers responding to similar requests at the Intermediate-Mid sub-level, which can be considered a positive indicator of his willingness to communicate. However, his speech is hampered by difficulties with conjugation and word forms, and also by some interjections of "that" and "I" which do not make sense in

the context of his speech. It appears that his narration attempt requires examiner intervention to bring it to a close.

Minimal Narration and Failed Narration at the Intermediate-Mid Level

In this section, I will treat one minimal narration and one failed narration sample from the Intermediate rating level. An Intermediate-Mid speaker produced a skeleton narration that was elicited by E1 approximately 20 minutes into the test. The transcript of this sub-sample follows, with spaces indicating redacted personal names and three question marks for unintelligible words. Words in bold in the English translation were spoken in English in the test:

ماذا فعلتَ في الصيف الماضي؟

في الصيف الماضي زرتُ ودرستُ في مصر

طيب حدثني كيف كانت الدراسة والزيارة؟

عفواً؟

كيف كانت الدراسة وكيف كانت الزيارة؟ كيف كانت؟

نعم كيف كيف كانت جيد مصر جميل جداً وسكنتُ في قاهرة*

طيب حدثني _____ أريد أن أعرف كيف كان اليوم الأول لكَ في مصر؟

ما شاء الله

ما شاء الله أوكاي يعني أنت ذهبت من أرميكا من جي أف كي أو من شيكاغو والطائرة وبعد ذلك أريد أن أعرف القصة

؟؟؟ خرجنا من مطار مطار

JFK

في مدينة نيويورك ووَقِفنا في بلاد* ألماني * ولكن بعد ذلك ركبنا طائرة لل مدينة قاهرة* وصلتُ وصلنا في قاهرة في* بعد الظهر وركبنا سيارة لل ركبنا سيارة للمركز حيث درسنا ونزلنا الأستاذ وتكلمنا مع هو* وبعد ذلك ذهبنا إلى

الأستاذ كان في المركز أو كان في الأوتوبيس؟

عفواً؟

الأستاذ كان في المركز أو كان في السيارة؟

Oh

كان في المركز نعم

وبعد ذلك؟

بعد ذلك ذهبنا إلى شقتنا ونعم بعد شقتنا ذهبنا مع الأستاذ إلى مطعم للعشاء*

وفي اليوم التالي؟

في يوم الثالث؟

اليوم التالي

Oh Oh

في اليوم التالي

Okay

استيقظنا في الساعة السابعة ونصف واستحممنا ومشينا إلى المركز لصف ودرسنا فصحى لساعتين وهنا كان هناك عطلة* لممكن خمسة عشر دقيقة وبعد ذلك درسنا عامية لساعة واحد*

What did you do last summer?

Last summer I visited and studied in Egypt

So tell me how was studying [there] and the trip?

Pardon?

How was the studying and how was the trip? How was it?

Yes how how was it? oh good Egypt is very beautiful and I lived in Cairo

Ok, tell me _____ how was your first day in Cairo?

My goodness!

My goodness, okay, you know, you went from America from JFK or Chicago and the plane and then after that I want to know the story

*??? we left from the airport the **JFK** airport in New York city we stopped in the German countries*after that we rode the plane to Cairo I arrived- we arrived in Cairo in the afternoon we rode in a car we rode in a car to the center where we studied and the teacher dropped us off and we talked to he* and after that we went to*

The teacher- was he at the center or was in the bus?

Pardon?

The teacher was in the center or was he in the bus?

***Oh** he was in the center yes*

And after that?

*After that we went to our apartment and yes after our apartment we went with the teacher to a restaurant for dinner**

And the following day?

On the third day?

The following day

***Oh oh** on the following day okay*

We woke up at 7:30 and we showered and walked to the center for class and we studied formal Arabic for two hours and there was a vacation for maybe 15 minutes and after that we studied dialect for one hour*

The examiner begins by asking what the test taker did last summer and then makes a more specific request for a story of the first day of the test taker's study abroad experience. The test taker begins a minimal narration by listing the events of his trip abroad, in which he produces clauses that are temporally related. He provides an orientation of place ("we left JFK airport") and actors ("we" and "the teacher"). However, the narration lacks evaluation and most importantly, reportable events, making this an attempt that can only be classified as minimal narration.

The incomplete nature of this narration is not only reflected in the content, but also in the examiner prompts about the other actor ("Was the teacher in the center or in the bus?") and requesting more information ("after that?" and "and the next day?"). The examiner's prompting is emblematic of the lacking nature of this attempted narration, and would not be necessary if the test taker had included reportable events, an evaluation of his experience or the experience of the other members of his group, or a coda to signal the narration's end. In terms of lexical errors, this test taker pronounces "dinner" in a clipped manner that makes it harder to interpret, and uses some forms of words that are incorrect ("countries" and "with he"). He also uses "vacation" when "break" would be more appropriate. However, these errors do not interfere with understanding the basic information he is communicating.

In contrast to the preceding minimal narration, an Intermediate-Mid test taker's attempt, approximately nine minutes into the interview, can be considered unsuccessful.

The following are the examiner prompts and the test taker responses, with bold indicating words that were uttered in English:

عندما كنت في تونس هل كان عندك حادثة أو قصة أو شيء يعني غريب حدث معك في تونس؟

احبتُ أن آكل اكلتُ كسكسي والبريك هو انا بيضة* فيه* والسمك ووشربتُ شاي قهوة وعصير ال البرتقال وكوكا معظم معظم من الشراب في ال الأمريكا

وهل سكنتَ مع طلاب في تونس؟
سكنتَ واحد* دقيقة من فضلك سكنتَ لا أعرف هذا* الكلمة آسف

طيب في شهر نومبر كان هناك عيد الشكر. ماذا فعلتَ في عطلة عيد الشكر في شهر نوفمبر؟

في الشهر في ال في هذا نوفمبر انا في كنتُ في البلد ال

Morocco

المغرب انا درستُ في المغرب كل هذا هذا الفصل وانا* انا وصلتُ من

Morocco

في في أمريكا أمس

yeah

لذلك أنا أتكلم معك

So

في عيد الشكر أنا وأصدقائي ذهبنا إلى مطعم _____ في مدينة _____ هو
مطعم أمريكي وأكلنا دجاج والطعام أميريكان* فيه

When you were in Tunisia, did you have an event or a story or something, you know, strange that happened to you in Tunisia?

I liked to eat I ate couscous and briik this is that an egg* in it and fish and I drank coffee and tea and orange juice and Coke most of the drinks in America*

And did you live with students in Tunisia?

"You lived"? One minute please "you lived" I don't know this word sorry*

Okay, in November, there was the Thanksgiving holiday. What did you do in the Thanksgiving break in the month of November?

*In the month in the this month of the November I am in I was in I arrived from **Morocco** Morocco I studied in Morocco all of this semester I arrived from **Morocco** in in America yesterday **yeah so** as a result I'm talking with you **so** on Thanksgiving my friends and I went to _____ restaurant in the city of _____ it's an American restaurant and we ate um chicken and American food there*

This test taker's first response does not constitute narration according to Labov's definition. The clauses are not temporally related so reversing the order of the actions would not affect understanding. The test taker reports what he ate and drank while abroad and expresses this in the past tense, but none of the information is unusual. The examiner attempts to expand the answer with a question about who the test taker lived with, but the test taker does not understand the word "lived" and the examiner subsequently abandons the question.

In the second attempt, the examiner asks the test taker what he did for Thanksgiving break. The test taker's second response forms a minimal narrative as the clauses are temporally linked, but again the test taker does not include reportable events. There are several lexical gaps or mispronunciations that interfere with communication in this exchange, especially the test taker's pronunciation of what sounds like "egg," his

failure to recognize the word "lived," and the switching he does between using the Arabic word for Morocco and the English word.

Superior Narrations

In contrast to the Intermediate sub-samples, both of the Superior rating level narration sub-samples were longer and more detailed. Both test takers spoke for approximately four minutes after the examiner finished his or her question. In that time, both test takers produced orientations, reportable events, and evaluated and concluded their narrations without enlisting examiner help. Their narrations also involved less first person language than many of the other narration responses at other rating levels.

Although I will not reproduce all the details here, a short summary of the stories should illustrate the more detailed nature of narrations at this level. The first story was from a young adult novel the test taker read. The main character in the book was a student who was attacked and raped by an older student at a party. The test taker says the character starts high school with this secret that she doesn't tell anyone until the end of the year. The test taker also states that she enjoyed the book because it showed problems her own students faced and also included how the girl used art to express her feelings. In addition to giving these details, this test taker also produced a pair of synonyms for the word "feelings" which is commonly considered a marker of good style in Arabic.

The second test taker told a story from his time in Morocco. A friend invited him to observe a religious ritual in a cave outside of a town she was studying. He described the story that drew the original Jewish inhabitants to the cave – concerning a rabbi who married a female devil – and how Muslims had continued the ritual of burning candles in this cave even after the Jewish community had left the area. The examiner asked him for

a description of the cave and he said that when he climbed to it and looked inside it was filled with people, cats, and "feelings of magic."

Although this data is limited and should be treated with some caution, it appears that stories such as these two Superior narrations required more specific vocabulary than test takers at Advanced or Intermediate rating levels produced or were able to retrieve and use appropriately in their narrations. Undoubtedly, a constellation of factors contributes to a test taker's ability to perform tasks like narration and description. However, it is also apparent that lexical richness must undergird these efforts to a large degree. This appears to be a promising area and one worthy of further exploration with larger numbers of samples.

Having presented my observations of the description and narration sub-samples, I will summarize and discuss the findings as well as the limitations of this research in chapter 5. I will also suggest in chapter 5 some directions for future research.

CHAPTER 5: CONCLUSIONS

In the previous chapter, I presented my findings from the full-length and sub-sample data. I addressed my first and second research questions by providing data on the words and words per minute for full-length tests from Advanced-Mid, Intermediate-Mid, and Superior rating levels; the TTRs for full-length tests at these levels; and the total tokens and TTRs for the description sub-samples that were longer than 100 tokens. I also generated combined word lists from the description sub-samples in an effort to isolate the shared vocabulary test takers used to respond to requests for description. I addressed my third research question by examining the frequency rankings for the shared vocabulary used in the description sub-samples and the full-length transcripts. For my fourth and final research question, I provided qualitative observations concerning the description and narration sub-samples.

In this chapter, I will summarize and discuss my findings and the study's limitations and provide reflections on directions for future research.

SUMMARY OF FINDINGS AND DISCUSSION

The findings for WPM and TTR of the full-length tests show that these measures can distinguish test takers at the Advanced-Mid rating level from the Intermediate-Mid level as represented in this data. This supports the assumption that Arabic is similar to other L2s and that, as is commonly understood, learners must be able to produce more words in order to surpass the Intermediate-Mid rating level. The TTR differences also indicate that the Advanced-Mid test takers produced more varied vocabulary than the Intermediate-Mid test takers. In other words, Advanced-Mid test takers were not simply producing more of the same words per minute but were instead drawing from a more

varied pool of words to respond to examiner questions. This suggests that curriculum designed for L2 learners of Arabic should specifically address the need to learn more varied words if students are to surpass the Intermediate-Mid rating level, rather than just aiming for faster flow of the same vocabulary. The TTR measurements also support the position that Arabic is similar to other languages, and that lexical resources—both their quantity and diversity—set undeniable boundaries on the quality of L2 Arabic speech.

However, the relationship between the Advanced-Mid rating level and the Superior rating level is less clear. The findings indicate that the Superior rating level test takers did not produce more words or a more varied group of words than the Advanced-Mid rating level test takers in the context of the OPI and the current data. This runs counter to common notions of improvement, which tend to include an underlying assumption of expanded and/or more varied lexical resources. I suggest here three possible explanations:

- First, it is possible that the words produced per minute and the variety of that vocabulary are simply not factors that distinguish the Advanced-Mid rating level from the Superior rating level. If this is the case, then further qualitative exploration comparing these two groups' speech may be needed to consider other possible differentiating factors.
- A second possibility is that Superior rating level test takers may possess larger and/or more varied lexical resources, but that the ACTFL OPI is not tapping that knowledge in a productive manner. This leaves open the possibility that either the test is tapping some of that knowledge in a receptive manner or that the test is not requiring this knowledge at all. If the latter is the case, then this will again lead back to the conclusion that vocabulary resources are not a deciding factor in the production of test

takers at this level. However, if the test is requiring the use of test takers' receptive vocabulary knowledge in some way, then further exploration of Arabic OPI examiner behavior and language is warranted.

- A third possibility is simply that the TTR is not a useful measure to use when comparing the lexical richness of these two groups or that the number of L2 test takers was insufficient to make subtle differences apparent. If this is the case, then a different measure may be needed or a larger pool of data may be necessary to reveal a finer distinction between the lexical richness of the Advanced-Mid and Superior test takers' speech.

The findings for shared vocabulary among Advanced-Mid, Superior, and Intermediate-Mid test takers showed that test takers use fairly similar shared words across all three groups, thus ruling out the assumption that shared vocabulary resources differentiate the Advanced-Mid test takers from those in the other two groups. Additionally, Superior rating level test takers used more words in common than Advanced-Mid and Intermediate-Mid test takers, which appears to run counter to my original assumption that Superior speakers' vocabularies would be more varied and therefore the shared word list would be smaller than those from the Advanced-Mid and Intermediate-Mid rating levels. Similarly, the findings regarding the frequency rankings for these words suggest that these L2 speakers are not producing less frequently used vocabulary in common, but are instead typically producing this shared vocabulary from among the 500 most commonly used words. This seems to imply that frequency rankings – as applied to L2 Arabic speakers' shared vocabulary – do not distinguish between L2 speakers' differing ability levels as measured in this test. This appears to conform to Bardel and Linqvist's findings with L2 speakers of French and Italian (Bardel & Lindqvist, 2011).

LIMITATIONS

The number of full-length tests that were transcribed and the number of description and narration attempts that were found in the data were necessarily limited. A larger number of test recordings might reveal differences between the Advanced-Mid rating level and the Superior rating level's lexical use that were not found in the present data; similarly, gathering data from L2 speakers responding to tests that have more rigid administration guidelines than the ACTFL OPI (as Margaret Malone did with the SOPI) might produce clearer distinctions. In addition to this, the data set represented a convenience sample and it is unclear what percentage this sample represented of the total number of Arabic tests administered in the same year. As a consequence, my ability to generalize from these findings is more limited than if I had been working with a truly random sample.

Another set of limitations was introduced by the nature of the test itself: the fact that examiners are not required to follow a set format means that test takers are asked questions in different ways and at different points in their tests. The lack of uniformity in examiner question delivery coupled with natural differences in examiner style could have affected the speech samples that were ultimately used in this study. There are also clearly differing expectations on the part of test takers. A small number of recordings included test takers asking questions about whether or not they could use formal Arabic or dialect; the responses these test takers received varied from examiner to examiner. There were also some recordings made over Skype or long-distance phone lines that may have affected the sound quality for both test takers and examiners; this introduced another potential source of variation unrelated to test takers' speaking abilities.

There are also some limitations to my qualitative observations about test taker speech due to the subjective nature of judging test takers' speech and their interactions with examiners. Some examiner intervention was fairly transparent, and I am confident that in some cases it reflected difficulty comprehending a word or words in the test taker's speech. However, other examiner interventions were more ambiguous, and it must be acknowledged that a lack of intervention does not necessarily imply that the examiner was able to understand the test taker's speech. In addition, the examiners' and my English ability may have introduced bias. Recent research has found support for the hypothesis that rater familiarity with test taker's L1—either as a first or second language—may bias their assessments of the test takers' L2 speaking ability (Carey, Mannell, & Dunn, 2011; Winke, Gass, & Myford, 2011; Xi & Mollaun, 2011). Although this is a minor concern given the exploratory nature of this research, it still may warrant further attention in future studies.

UNDERSTANDING ADVANCED

My working hypotheses for this research were: 1) that the Advanced rating level's word production would be higher than the Intermediate rating level's and lower than the Superior rating level's, 2) that lexical richness would distinguish higher rating levels, and 3) that task type may affect the vocabulary L2 Arabic speakers produced in this speaking test. The first hypothesis was only partially supported: Advanced rating level test takers' word production distinguished them from Intermediate-Mid rating level speakers but not from the Superior rating level test takers. The second hypothesis was also only partially supported: the TTR lexical richness measure distinguished the Advanced rating level

from the Intermediate rating level test takers but did not distinguish the Advanced rating level from the Superior rating level test takers. The third assumption was explored in the sub-samples of description and narration. Although anecdotally the quality of the vocabulary appeared to vary among the rating levels, these apparent differences were not reflected in the measures chosen, and I did not find increased word production or TTRs in the Advanced or Superior rating levels.

The significant differences in WPM and TTR between the Advanced-Mid and Intermediate-Mid rating levels lend empirical support to Al-Batal's position that vocabulary may be regarded as a defining obstacle, obstructing L2 Arabic learners' paths to the Advanced rating levels. While this may appear disheartening at first, it is in fact an encouraging finding. This means that—at least in this respect—Arabic is like other foreign languages, and this commonality could be beneficial to Arabic language researchers, instructors and learners. For researchers, it implies that vocabulary in Arabic could be a way in which L2 learners can aspire to levels similar to those of L1 speakers. At the very least, this indicates that vocabulary is likely to be as useful a vein of research-based findings in Arabic as it has been in other languages. For Arabic language instructors and curriculum designers, it encourages a continued focus on vocabulary expansion—particularly one that regularly requires productive control from learners. For students of the language, it makes clear that an investment in an expansive and diverse vocabulary pool is a necessary step that all L2 Arabic learners should strive to take. It is one that will pay off if they push themselves to use their vocabulary knowledge in a productive manner.

Turning to the qualitative observations, they represent an important contribution to documenting what Arabic L2 learners of varying abilities can accomplish when asked to describe a city or tell a story. Previously, samples of this kind were heard only by

examiners in the course of administering these language tests or, more recently, through annotations and exemplars of the rating levels available as a result of the National Arabic Consensus project, begun in 2009. The NSEP Program funded this project as part of its efforts to make the Language Flagship programs' innovations available to all interested Arabic testers and learners, and these are available online[31].

However, the qualitative observations raise more questions than they answer. One of the largest questions is: How parsimonious can a narration or description be and still be considered to have sufficiently met the requirements of the task? It appears to be easier to define the necessary elements of narration, among them the primary need for a "reportable" event. If a reportable event is at the core of the requirements for narration, then perhaps the other elements of an orientation and evaluation can be considered of minimal importance. However, it is difficult to state with precision what "minimal" will mean in a testing context, particularly in regard to description. In addition, I hypothesized that the descriptions of Advanced rating level test takers would distinguish the city being described from other cities in the world. This was often not the case. It appears that this distinguishing description was more often accomplished at the Superior rating level and only sporadically found at the Advanced rating levels.

DIRECTIONS FOR FUTURE RESEARCH

I will suggest here three potential future directions for this research. The first would further explore the current data and the use of Buckwalter and Parkinson's frequency rankings. The second would focus on teacher judgments of vocabulary

[31] The ACTFL Arabic language annotations and examples can be found at: http://actflproficiencyguidelines2012.org/arabic/index.php.

frequency rankings that could be applied to current or future data. The third would involve more qualitative exploration of the current data.

In this study, I compared the shared vocabulary pools of different rating levels. It was clear from this data that Advanced-Mid test takers did not collectively produce a pool of vocabulary that distinguished them from Superior and Intermediate-Mid rating level test takers. However, this does not indicate anything concerning the frequency rankings for words that test takers used as part of their individual vocabulary pools. A useful future direction of research would be to compare the frequency rankings of individual test takers' vocabulary to one another to see if frequency rankings of Advanced-Mid test takers' individual vocabulary use cluster in a different frequency range than the vocabulary of Intermediate-Mid or Superior test takers. If this were the case, then it would challenge the hypothesis that frequency rankings could not distinguish between these groups of Arabic L2 speakers. If no pattern were detected, then it would lend support to the position that Buckwalter and Parkinson's frequency rankings may not be a useful measure for examining the speech of these groups of L2 speakers. This could indicate a similarity in the individual vocabulary pools or it could suggest that Buckwalter and Parkinson's frequency rankings are not an appropriate measure for vocabulary produced in Arabic L2 speech. The fact that words like "Arabic" and "I study" were ranked beyond the 1,000 most commonly used words make it appear that at least some words may need to be re-evaluated by those with regular contact with Arabic L2 learners.

Teacher judgments (TJs) could be used as an alternative method to Buckwalter and Parkinson's rankings or in combination with them. Camilla Bardel and her colleagues developed and evaluated the usefulness of a lexical profiler for Swedish L1 speakers of French and Italian by incorporating TJs (Bardel, Gudmundson, & Lindqvist,

2012). They hypothesized that including TJs in lexical profiles would improve their ability to gauge learners' lexical ability if teachers could reliably identify which words should be considered thematic vocabulary or cognates for L2 learners (Bardel, Gudmundson, Lindqvist, 2012, p 270). After incorporating TJs into a second version of their lexical profiler, the new profiler provided clearer distinctions between learners and native-speakers than one based on frequency data alone. A similar method could be employed with teachers of Arabic. Assuming that internally consistent TJs could be gathered from Arabic teachers, these frequency judgments could be used to generate rankings for vocabulary. The rankings could then be used to profile L2 speakers in an effort to see if the TJs correlated with speaking test ratings.

A third direction for further research would involve expanding the qualitative observations provided in the present study. My focus was on offering observations about the Advanced rating levels' collective abilities and, as a result, less time was spent examining failed narrations in particular. This could be an enlightening research line to consider. Richard Robin has analyzed narration found in 54 OPIs conducted with L2 speakers of Russian and reported that failed narration among these test takers occurred more frequently at the Intermediate-High and Advanced-Low levels, which he interpreted as an indication that these learners were less savvy in their avoidance strategies than learners at other levels (Robin, 2011). This is worth exploring in Arabic language-learner data. First, Robin's findings appear to support Liskin-Gasparro's dissertation data findings with L2 speakers of Spanish, i.e. that Advanced learners were able to provide more detailed narratives than Intermediate learners. If similar differences are found in L2 Arabic speakers' data, then the ability to narrate effectively—by producing language that simultaneously advances a story while also avoiding individual learner's linguistic pitfalls—could differentiate between Advanced and Intermediate rating levels.

This third avenue might also be a productive one to pursue given the fact that the quantitative differences in this study did not serve to differentiate between the Advanced and the Superior rating levels in particular. As I noted in the limitations, the TTR measure may not have been sufficiently sensitive to capture variation between these two rating groups, there may have been too limited a number of samples in the Superior rating group, or the data set may have included too much variation in examiner behavior to provide an accurate picture. However, the lack of a significant difference between the Advanced and Superior rating level groups' words per minute coupled with the apparent qualitative difference between both groups' speech samples is intriguing. To my mind, this suggests that rather than Superior rating level speakers simply increasing the number of words they produce, they might instead be demonstrating a kind of linguistic consolidation occurring at this stage. It is possible that Superior rating level test takers are able to make more effective use of their lexical resources in order to accomplish the test tasks using fewer words. If this is the case, then the Advanced rating level may be the point at which the word production growth levels off; perhaps Superior rating level L2 speakers are approaching a threshold at which they may be both more parsimonious and more communicative[32].

Finally, since the ACTFL Advanced rating level is the functional level that most learners aspire to and the one they are most likely to reach in traditional instructional settings, I hope this work will be seen as a contribution to an area that is rich in its

[32] As I mentioned in the literature review, Malone's findings with SOPI data and Read and Nation's findings using IELTS data both suggest that as L2 speakers improve their ability to produce more words and more varied words in their speech tends to increase. The data from the present study supports that finding, at least when examining the difference between the Intermediate and Advanced rating level test takers. While more research is of course warranted, this lends some credence to the validity claims of ACTFL OPI users because it demonstrates the test's ability to discriminate between Advanced and Intermediate rating levels in a LCTL.

potential for further research and practice-oriented explorations. To that end, I will suggest one more area worthy of attention on a discipline-wide level: an empirically grounded consideration of how the ACTFL OPI is administered and what types of L2 speaking performance are elicited in LCTLs like Arabic. It seems likely that the OPI will remain a widely used assessment tool in the U.S., particularly among LCTLs, as changes have often been slower to reach them for various reasons. The intersection of candidate and examiner behavior seems to me to be a clear starting point in investigating the ways in which speaking tests like the OPI might be modified to capture a clearer and more accurate understanding of L2 speaker ability.

Appendix A: Advanced Rating Level Shared Vocabulary across half or more of 12 City Descriptions

Table A-1: Advanced shared vocabulary produced in half or more of 12 city description sub-samples

English	Arabic	Number of sub-samples	Buckwalter & Parkinson's frequency number [or closest equivalent]
And	و	12	2
The city	المدينة	11	144
In/at	في	11	3
City	مدينة	11	144
From	من	11	4
There	هناك	10	77
So, thus	ف	9	21
It/that	ما	9	28
Like, similar to	مثل	9	86
Like, you know	يعني	9	751
To	إلى	8	9
With	مع	8	17
This (masc.)	هذا	8	16
But	ولكن	8	91
To/that	أن	7	13
Or	أو	7	23
Beautiful (fem.)	جميلة	7	304
A lot (masc.)	كثير	7	55
All	كل	7	19
Not	ليس	7	59
Yes	نعم	7	22
She/it	هي	7	33
The people	الناس	6	[انسان - 204]
according to	بالنسبة	6	[نسبة - 155]
That (masc.)	ذلك	6	36
Thing	شيء	6	39
Much (fem.)	كثيرة	6	55
This (fem.)	هذه	6	22

Appendix B: Intermediate Rating Level Shared Words across 19 City Descriptions

Table B-1: Intermediate shared vocabulary produced in half or more of 19 city description sub-samples

English	Arabic	Number of sub-samples	Buckwalter & Parkinson's frequency number [or closest equivalent]
And	و	19	2
In/at	في	17	3
From	من	16	4
City	مدينة	15	244
Yes	نعم	14	22
The city	المدينة	11	244
The people	الناس	10	[انسان - 204]
All	كل	10	19
This (masc.)	هذا	10	16
But	ولكن	10	91
How	كيف	9	67

Appendix C: Shared Vocabulary Words from Test Takers Describing the Same City

Table C-1: Shared vocabulary between two Advanced-Mid rating level descriptions of the same city

English	Arabic	Number of sub-samples	Buckwalter & Parkinson's frequency number [or closest equivalent]
Culture	الثقافة	2	519
Syrians	السوريين	2	578
The city	المدينة	2	144
The people	الناس	2	[انسان - 204]
Pretty (fem.)	جميلة	2	304
Damascus	دمشق	2	N/A
Syria	سوريا	2	519
In/at	في	2	3
I was	كنتُ	2	[كان - 10]
Because	لأنه	2	57
A city	مدينة	2	144
With	مع	2	17
From	مِن	2	4
Yes	نعم	2	22
There	هناك	2	77
And	و	2	2

Table C-2: Shared vocabulary between three Superior and Advanced-High rating level descriptions of the same city

English	Arabic	Number of sub-samples	Buckwalter & Parkinson's frequency number [or closest equivalent]
Alexandria	الاسكندرية	3	N/A
The sea	البحر	3	507
The city	المدينة	3	144
To	إلى	3	9
Very	جداً	3	N/A
In/at	في	3	3
A city	مدينة	3	144
From	مِن	3	4
There	هناك	3	77
And	و	3	2
But	ولكن	3	91
Like, you know	يعني	3	751

Table C-3: Shared vocabulary between two Intermediate-High rating level descriptions of the same city

English	Arabic	Number of sub-samples	Buckwalter & Parkinson's frequency number [or closest equivalent]
The city	المدينة	2	144
History	تاريخ	2	N/A
Fez	فاس	2	N/A
In/at	في	2	3
Old (masc.)	قديم	2	499
Old (fem.)	قديمة	2	499
Like, similar to	مثل	2	86
For example	مثلا	2	[مثل - 86]
A city	مدينة	2	144
Moroccan	مغربي	2	N/A
Place	مكان	2	179
From	من	2	4
This (masc.)	هذا	2	16
Here	هنا	2	159
There	هناك	2	77
And	و	2	2
There is	يوجد	2	139

REFERENCES

Abdalla, M. (2006). Arabic immersion and summer programs in the United States. In K. Wahba, Z. A. Taha & L. England (Eds.), *Handbook for Arabic Language Teaching Professionals in the 21st Century* (pp. 317-330). Mahway, New Jersey: Lawrence Erlbaum Associations.

Al-Batal, M. (1995). Issues in the teaching of the productive skills in Arabic. In M. Al-Batal (Ed.), *The teaching of Arabic as a foreign language: Issues and directions* (pp. 115-133): American Association of Teachers of Arabic.

Al-Batal, M. (2006). Playing with words: Teaching vocabulary in the Arabic curriculum. In K. M. Wahba, Z. A. Taha & L. England (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (pp. 331-340). Mahwah, N.J.: Lawrence Erlbaum Associates.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Bardel, C., Gudmundson, A., & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production. *Studies in Second Language Acquisition, 34*(Special Issue 02), 269-290. doi: 10.1017/S0272263112000058

Bardel, C., & Lindqvist, C. (2011). Developing a lexical profiler for spoken French L2 and Italian L2: The role of frequency, thematic vocabulary and cognates. [Article]. *EUROSLA Yearbook, 11*(1), 75-93. doi: 10.1075/eurosla.11.06bar

Belnap, R. K. (1987). Who's taking Arabic and what on earth for? A survey of students in Arabic language programs. *Al- cArabiyya, 20*(1&2), 29-42.

Bergman, E. M. (2009). Arabic: Meeting the challenges. *Journal of the National Council of Less Commonly Taught Languages, 6*, 1-13.

Boudelaa, S., & Marslen-Wilson, W. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods, 42*(2), 481-487. doi: 10.3758/brm.42.2.481

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25. doi: 10.1191/0265532203lt242oa

Brown, A. V. (2009). Less commonly taught language and commonly taught language students: A demographic and academic comparison. *Foreign Language Annals, 42*(3), 405-423. doi: 10.1111/j.1944-9720.2009.01036.x

Buckwalter, T., & Parkinson, D. (2011). *A frequency dictionary of Arabic: Core vocabulary for learners*. New York, NY: Routledge.

Bulté, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time: the case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies, 18*(Special Issue 03), 277-298. doi: 10.1017/S0959269508003451

Byrnes, H. (2005). The privilege of the less commonly taught languages: Linking literacy and advanced L2 capacities. *Journal of the National Council of Less Commonly Taught Languages*.

Byrnes, H., Weger-Guntharp, H., & Sprang, K. A. (2006). Locating the advanced learner in theory, research, and educational practice: An introduction. *Educating for advanced foreign language capacities: Constructs, curriculum, instruction, assessment* (pp. 1-14). Washington, D.C.: Georgetown University Press.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? [Article]. *Language Testing, 28*(2), 201-219. doi: 10.1177/0265532210393704

Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals, 36*(4), 498-506. doi: 10.1111/j.1944-9720.2003.tb02139.x

Chambless, K. S. (2012). Teachers' oral proficiency in the target language: Research on its role in language teaching and learning. *Foreign Language Annals, 45*(s1), s141-s162. doi: 10.1111/j.1944-9720.2012.01183.x

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13. doi: 10.1111/j.1745-3992.2009.00165.x

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly, 33*(2), 185–209. doi: 10.2307/3587717

Council of Europe. (2001). The common european framework of reference for languages. http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp
accessed on May 21, 2012.

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*(3), 573-605. doi: 10.1111/j.1467-9922.2010.00568.x

Crossley, S., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What Is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45*(1), 182-193. doi: 10.5054/tq.2010.244019

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, Milton, J and Treffers-Daller, J (Ed.), *Modelling and Assessing Vocabulary Knowledge*.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23*(1), 11-22. doi: 10.1111/j.1944-9720.1990.tb00330.x

David, A. (2008). A developmental perspective on productive lexical knowledge in L2 oral interlanguage. *Journal of French Language Studies, 18*(Special Issue 03), 315-331. doi: 10.1017/S0959269508003475

Eisele, J. (2006). Developing frames of reference for assessment and curricular design in a diglossic L2: From skills to tasks (and back again). In K. M. Wahba, Z. A. Taha & L. England (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (pp. xxxiii, 477 p.). Mahwah, N.J.: L. Erlbaum Associates.

Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics, 16*(4), 409-441. doi: 10.1093/applin/16.4.409
Ellis, R. (2013). Changing trends in language teaching research. *Language Teaching Research, 17*(2), 141-143. doi: 10.1177/1362168812460807
Fakhri, A. (1984). The use of communicative strategies in narrative discourse: A case study of Moroccan Arabic as a second language. *Language Learning, 34*(3), 15-37.
Fedchak, K. D. (2007). *An empirical investigation of Russian interlanguage at the Superior level and the perspective of the educated native speaker*. Ph.D. 3265669, Bryn Mawr College, United States -- Pennsylvania. Retrieved from http://ezproxy.lib.utexas.edu/login?url=http://search.proquest.com/docview/304899271?accountid=7118
http://findit.lib.utexas.edu/utaustin?genre=article&sid=ProQ:&atitle=&title=An+empirical+investigation+of+Russian+interlanguage+at+the+Superior+level+and+the+perspective+of+the+educated+native+speaker&issn=&date=2007-01-01&volume=&issue=&spage=&au=Fedchak%2C+Kimberly+D. ProQuest Dissertations & Theses (PQDT) database.
Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*(02), 275-301. doi: 10.1017/S0272263104262064
Fulcher, G. (1996). Invalidating validity claims for the ACTFL oral rating scale. *System, 24*(2), 163-172. doi: 10.1016/0346-251x(96)00001-2
Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344. doi: 10.1191/0265532203lt259oa
George, M. G. (2011). *Teacher scaffolding of oral language production*. 3449456 Ph.D., The University of Arizona, United States -- Arizona. Retrieved from http://proxygw.wrlc.org/login?url=http://search.proquest.com/docview/862553152?accountid=11243 ProQuest Dissertations & Theses Full Text database.
Glisan, E. W., Swender, E., & Surface, E. A. (2013). Oral proficiency standards and foreign language teacher candidates: Current findings and future research directions. *Foreign Language Annals*, n/a-n/a. doi: 10.1111/flan.12030
Halleck, G. B. (1992). The oral proficiency interview: Discrete point test or a measure of communicative language ability? *Foreign Language Annals, 25*(3), 227–231. doi: 10.1111/j.1944-9720.1992.tb00532.x
Halleck, G. B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. [Article]. *Modern Language Journal, 79*(2), 223.
Halleck, G. B. (1996). Interrater reliability of the OPI: Using academic trainee raters. *Foreign Language Annals, 29*(2), 223-233. doi: 10.1111/j.1944-9720.1996.tb02329.x
Harlow, L. L., & Muyskens, J. A. (1994). Priorities for intermediate-level language instruction. *The Modern Language Journal, 78*(2), 141-154.

Hellman, A. B. (2011). Vocabulary size and depth of word knowledge in adult-onset second language acquisition. [Article]. *International Journal of Applied Linguistics, 21*(2), 162-182. doi: 10.1111/j.1473-4192.2010.00265.x

Henning, G. (1992). The ACTFL oral proficiency interview: Validity evidence. *System, 20*(3), 365-372.

Herzog, M. (2003). Impact of the proficiency scale and the oral proficiency interview on the foreign language program at the Defense Language Institute Foreign Language Center. *Foreign Language Annals, 36*(4), 566-571. doi: 10.1111/j.1944-9720.2003.tb02146.x

Husseinali, G. (2006). Who is studying Arabic and why? A survey of Arabic students' orientations at a major university. *Foreign Language Annals, 39*(3), 395-412.

Ilieva, G. N. (2012). Hindi heritage language learners' performance during OPIs: Characteristics and pedagogical implications. *Heritage Language Journal, 9*(2), 18-36.

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of a successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition, 16*, 73-98.

Isurin, L. (2012). Superior speakers or 'super' Russian: OPI guidelines revisited. In V. Makarova (Ed.), *Russian language studies in North America: New perspectives from theoretical and applied linguistics*: Anthem Press.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 24-49.

Johnson, M. (2000). Interaction in the oral proficiency interview: Problems of validity. *Pragmatics, 10*(2), 215–231.

Johnson, M. (2001). *The art of nonconversation : a reexamination of the validity of the oral proficiency interview*. New Haven: Yale University Press.

Kagan, O., & Friedman, D. (2003). Using the OPI to place heritage speakers of Russian. *Foreign Language Annals, 36*(4), 536-545. doi: 10.1111/j.1944-9720.2003.tb02143.x

Khaldieh, S. A. (2001). The relationship between knowledge of icraab, lexical knowledge, and reading comprehension of nonnative readers of Arabic. [Feature Article]. *Modern Language Journal, 85*(3), 416-431. doi: 10.1111/0026-7902.00117

Khoury, G. (2008). *Vocabulary acquisition in Arabic as a foreign language: The root and pattern strategy*.

Labov, W. (1972). *Language in the inner city: Studies in the black English vernacular*. Philadelphia: University of Pennsylvania Press.

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System, 20*(3), 373-386. doi: 10.1016/0346-251x(92)90047-7

Lindqvist, C. (2010). La richesse lexicale dans la production orale de l'apprenant avancé de français. [Article]. *Canadian Modern Language Review, 66*(3), 393-420. doi: 10.3138/cmlr.66.3.393

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. [Article]. *IRAL: International Review of Applied Linguistics in Language Teaching, 49*(3), 221-240. doi: 10.1515/iral.2011.013

Liskin-Gasparro, J. E. (1993). *Talking about the past: An analysis of the discourse of Intermediate High and Advanced level speakers of Spanish*. Ph.D. 9413541, The University of Texas at Austin, United States -- Texas. Retrieved from http://ezproxy.lib.utexas.edu/login?url=http://search.proquest.com/docview/304069097?accountid=7118

http://findit.lib.utexas.edu/utaustin?genre=article&sid=ProQ:&atitle=&title=Talking+about+the+past%3A+An+analysis+of+the+discourse+of+Intermediate+High+and+Advanced+level+speakers+of+Spanish&issn=&date=1993-01-01&volume=&issue=&spage=&au=Liskin-Gasparro%2C+Judith+Elaine Dissertations & Theses @ University of Texas - Austin; ProQuest Dissertations & Theses (PQDT) database.

Liskin-Gasparro, J. E. (1996a). Narrative strategies: A case study of developing storytelling skills by a learner of Spanish. [Article]. *Modern Language Journal, 80*(3), 271.

Liskin-Gasparro, J. E. (1996b). Circumlocution, communication strategies, and the ACTFL Proficiency Guidelines: An analysis of student discourse. *Foreign Language Annals, 29*(3), 317-330. doi: 10.1111/j.1944-9720.1996.tb01245.x

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching, 39*(03), 167–190. doi: 10.1017/S0261444806003557

Lorenzo-Dus, N., & Meara, P. (2005). Examiner support strategies and test-taker vocabulary. *International Review of Applied Linguistics in Language Teaching (IRAL), 43*(3), 239-258.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*(4), 415-437. doi: 10.1191/0265532205lt303oa

Macaro, E. (2003). *Teaching and learning a second language: a review of recent research*. London ; New York: Continuum.

Malone, M. E. (1999). *The development of the English speaking test: An investigation of reliability and validity*. Ph.D. 9955575, Georgetown University, United States -- District of Columbia. Retrieved from http://proquest.umi.com/pqdweb?did=730647081&Fmt=7&clientId=48776&RQT=309&VName=PQD

Meredith, R. A. (1990). The Oral Proficiency Interview in real life: Sharpening the scale. *The Modern Language Journal, 74*(3), 288-296.

Mikhailova, J. (2007a). Rethinking description in the Russian SOPI: Shortcomings of the Simulated Oral Proficiency Interview. *Foreign Language Annals, 40*(4), 584-603. doi: 10.1111/j.1944-9720.2007.tb02882.x

Mikhailova, J. (2007b). Lexical complexity of learner discourse: Interpersonal and presentational mode descriptions in Russian. *Russian Language Journal, 57*.
Mikhailova, J. V. (2005). *Comparison of interpersonal and presentational description in Russian oral proficiency testing*. Ph.D. 3176428, The Ohio State University, United States -- Ohio. Retrieved from http://proquest.umi.com/pqdweb?did=921029761&Fmt=7&clientId=48776&RQT=309&VName=PQD
Milton, J. (2008). French vocabulary breadth among learners in the British school and university system: comparing knowledge over time. *Journal of French Language Studies, 18*(Special Issue 03), 333-348. doi: 10.1017/S0959269508003487
Milton, J. (2010). The development of vocabulary breadth across the CEFR levels: A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. In I. Bartning, M. Martin and I. Vedder (Ed.), *Second Language Acquisition and Testing in Europe* (pp. 211-232).
Milton, J., & Alexiou, T. (2009). Vocabulary size and the Common European Framework of Reference for Languages. In B. Richards, M. H. Daller, D. Malvern, P. Meara, J. Milton & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 194-211).
Morkus, N. (2009). *The realization of the speech act of refusal in Egyptian Arabic by American learners of Arabic as a foreign language*. Ph.D. 3420606, University of South Florida, United States -- Florida. Retrieved from http://ezproxy.lib.utexas.edu/login?url=http://search.proquest.com/docview/750372366?accountid=7118
http://findit.lib.utexas.edu/utaustin?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:ProQuest+Dissertations+%26+Theses+%28PQDT%29&atitle=&title=The+realization+of+the+speech+act+of+refusal+in+Egyptian+Arabic+by+American+learners+of+Arabic+as+a+foreign+language&issn=&date=2009-01-01&volume=&issue=&spage=&au=Morkus%2C+Nader&isbn=9781124187761&jtitle=&btitle= ProQuest Dissertations & Theses (PQDT) database.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York City: Newbury House Publishers.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge ; New York: Cambridge University Press.
National Middle East Language Resource Center. (2011). Middle East language learning in U.S. higher education: Ten years after 9/11.
National Security Education Program. (2012). National Security Education Program annual report.
Norris, J. M., & Pfeiffer, P. C. (2003). Exploring the uses and usefulness of ACTFL oral proficiency ratings and standards in college foreign language departments. *Foreign Language Annals, 36*(4), 572–581. doi: 10.1111/j.1944-9720.2003.tb02147.x

O'Loughlin, K. (2000). The impact of gender in the IELTS oral interview. In R. Tulloh (Ed.), *IELTS Research Reports, Volume 3* (pp. 1-28). Canberra: IELTS Australia.
Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales mesure fiable du niveau de competence langagière. (French). [Article]. *Canadian Modern Language Review, 63*(1), 107-125.
Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports, 6*, 207-231.
Richards, B., Daller, M. H., Malvern, D. D., Meara, P., Milton, J., & Treffers-Daller, J. (Eds.). (2009). *Vocabulary studies in first and second langauge acquisition: The interface between theory and application*. Hampshire, England: Palgrave MacMlllan.
Rifkin, B. (2002). A case study of the acquisition of narration in Russian: At the intersection of foreign language education, applied linguistics, and second language acquisition. *The Slavic and East European Journal, 46*(3), 465-481.
Rifkin, B. (2003). Oral proficiency learning outcomes and curricular design. *Foreign Language Annals, 36*(4), 582-588. doi: 10.1111/j.1944-9720.2003.tb02148.x
Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. [Article]. *Modern Language Journal, 89*(1), 3–18. doi: 10.1111/j.0026-7902.2005.00262.x
Rivera, G. M., & Matsuzawa, C. (2007). Multiple-Language program assessment: Learners' perspectives on first- and second-year college foreign language programs and their implications for program improvement. *Foreign Language Annals, 40*(4), 569-583. doi: 10.1111/j.1944-9720.2007.tb02881.x
Robin, R. M. (2011). Narration and narrative in L2 speakers of Russian. *Foreign Language Annals, 44*(1), 153-180. doi: 10.1111/j.1944-9720.2011.01121.x
Robin, R. M. (2012). Lexicalized aspectual usage in oral proficiency interviews. *The Modern Language Journal, 96*(1), 34-50. doi: 10.1111/j.1540-4781.2012.01292.x
Ross, S. (1996). Formulae and inter-interviewer variation in oral proficiency interview discourse. *Prospect: the Journal of the Adult Migrant Education Program, 11*(3), 3-16.
Ryding, K. (2006). Teaching Arabic in the United States. In K. Wahba, Z. A. Taha & L. England (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (pp. 13-20). Mahwah, New Jersey: Lawrence Erlbaum Associates.
Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research, 27*(3), 343-360. doi: 10.1177/0267658310395851
Schmitt, N. (2010). *Research vocabulary: A vocabulary research manual*.
Scott, M. (2012). WordSmith Tools version 6. Liverpool: Lexical Analysis Software.
Shohamy, E. (1983). Rater reliability of the Oral Interview speaking test. *Foreign Language Annals, 16*(3), 219-222. doi: 10.1111/j.1944-9720.1983.tb01456.x
Soliman, I. A. (2012). Center for Arabic Study Abroad full year program annual report 2011-2012.

Surface, E. A., & Dierdorff, E. C. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals, 36*(4), 507-519. doi: 10.1111/j.1944-9720.2003.tb02140.x

Swender, E. (1999). ACTFL Oral Proficiency Interview tester training manual. Yonkers, NY: American Council on the Teaching of Foreign Languages.

Swender, E. (2003). Oral proficiency testing in the real world: Answers to frequently asked questions. *Foreign Language Annals, 36*(4), 520-526. doi: 10.1111/j.1944-9720.2003.tb02141.x

Taguchi, N. (2007). Task difficulty in oral speech act production. *Applied Linguistics, 28*(1), 113-135. doi: 10.1093/applin/aml051

Thompson, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals, 28*(3), 407-422. doi: 10.1111/j.1944-9720.1995.tb00808.x

Vivrette, J. (2010). Cultivating awareness: Register and context in first-year Arabic Retrieved April 21, 2012, from http://blc.berkeley.edu/index.php/blc/post/cultivating_awareness_register_and_context_in_first-year_arabic/

Walker, J. L. (1973). Opinions of university students about language teaching. *Foreign Language Annals, 7*(1), 102-105.

Watanabe, S. (2003). Cohesion and coherence strategies in paragraph-length and extended discourse in Japanese oral proficiency interviews. *Foreign Language Annals, 36*(4), 555-565. doi: 10.1111/j.1944-9720.2003.tb02145.x

Winke, P., & Aquil, R. (2006). Issues in developing standardized tests of Arabic proficiency. In K. M. Wahba, L. England & Z. A. Taha (Eds.), *A Handbook for Arabic Language Teaching Professionals in the 21st Century* (pp. 221-235): Mahwah, NJ: Lawrence Erlbaum Associates.

Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples *TOEFL iBT® Research Report*: ETS.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255. doi: 10.1111/j.1467-9922.2011.00667.x

Zareva, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System, 33*(4), 547-562. doi: 10.1016/j.system.2005.03.005

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? [Article]. *Second Language Research, 23*(2), 123-153.